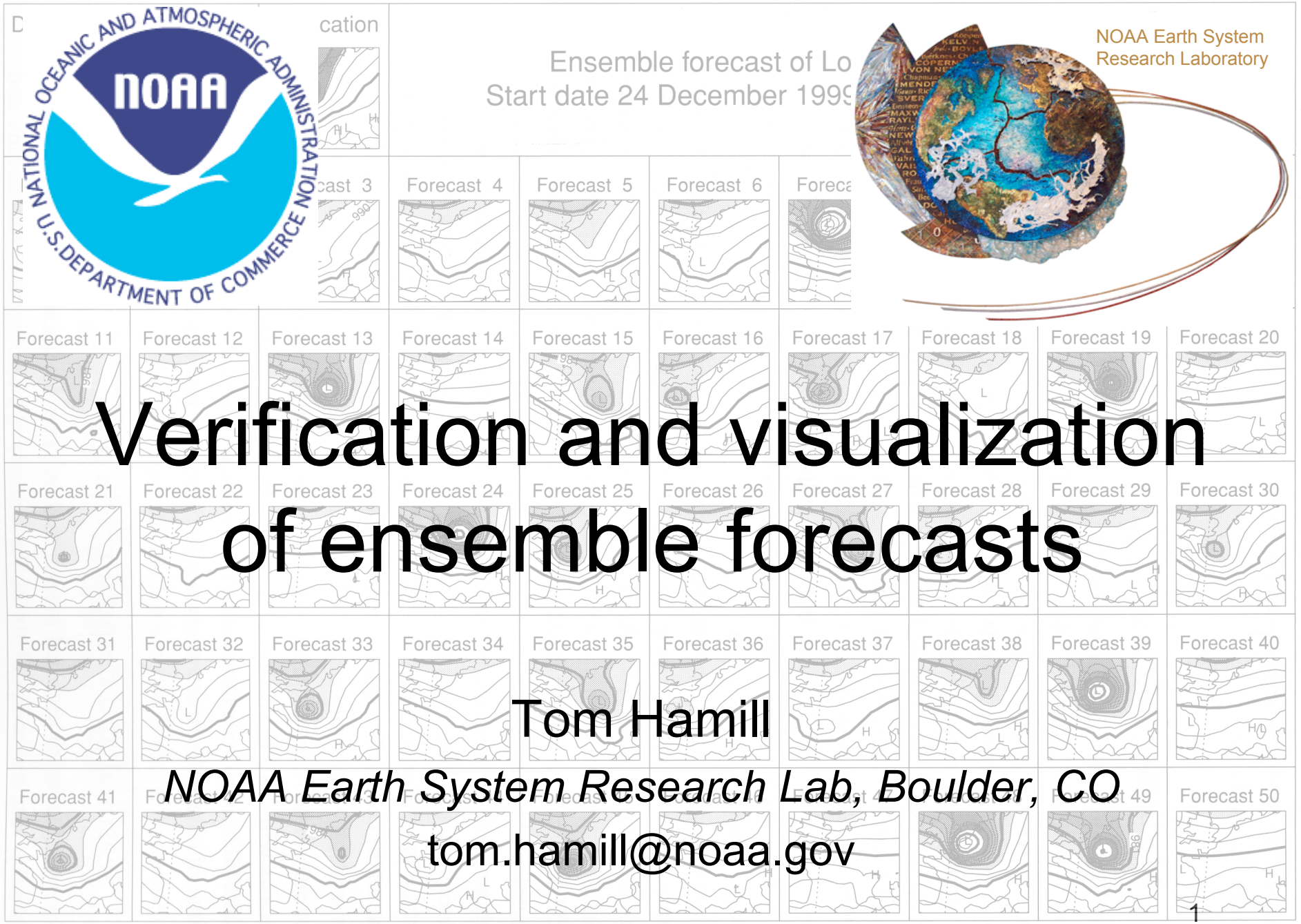


NOAA Earth System Research Laboratory

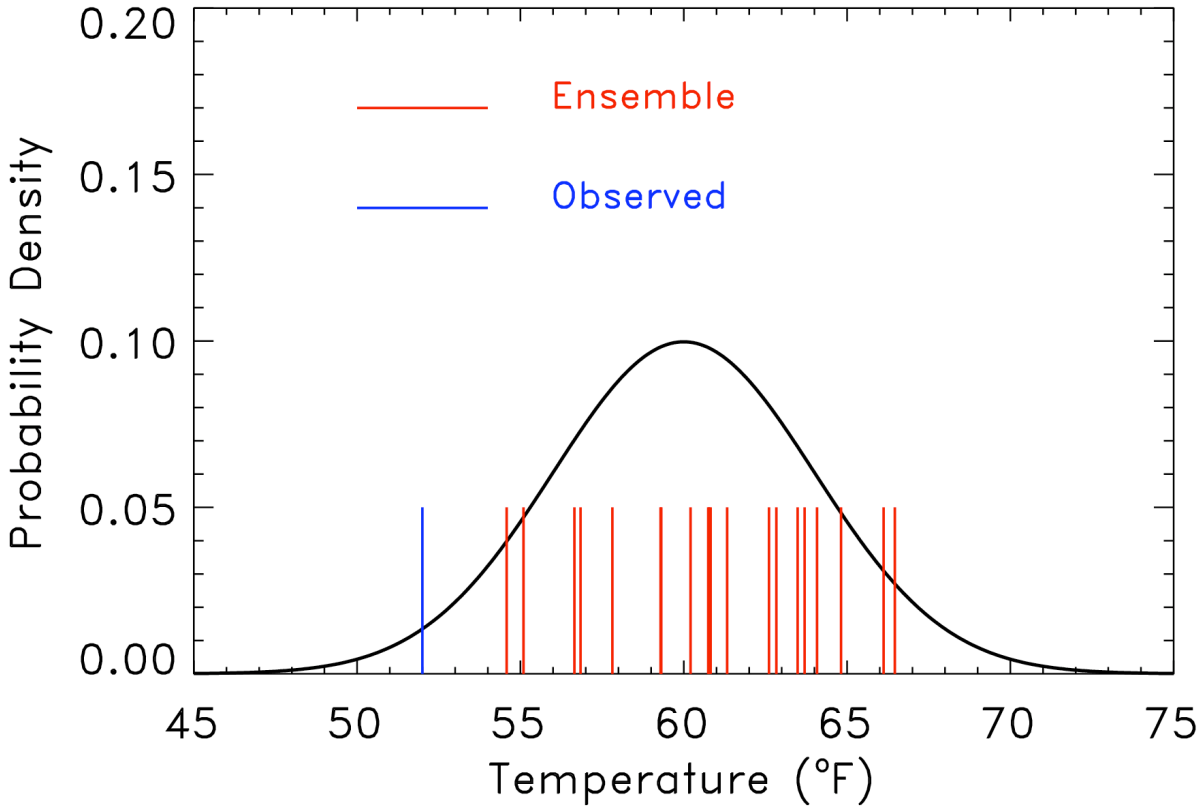
Verification and visualization of ensemble forecasts

Tom Hamill

NOAA Earth System Research Lab, Boulder, CO

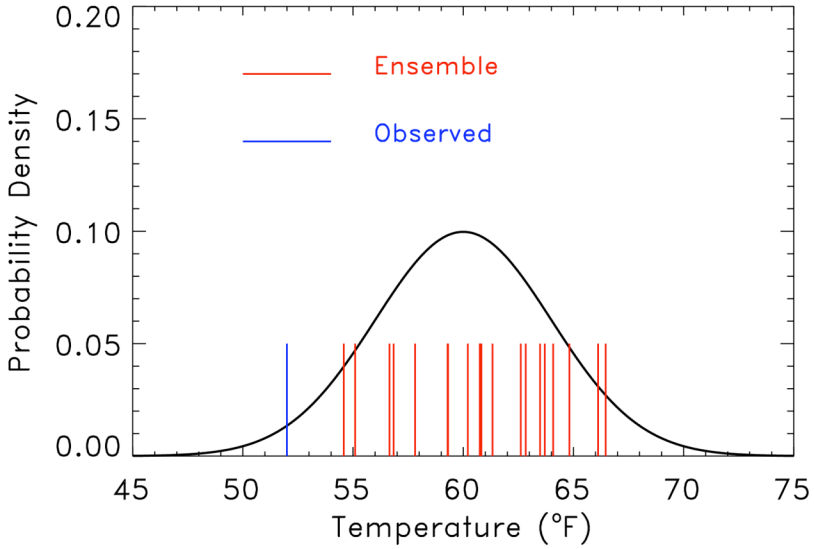
tom.hamill@noaa.gov

What constitutes a “good” ensemble forecast?

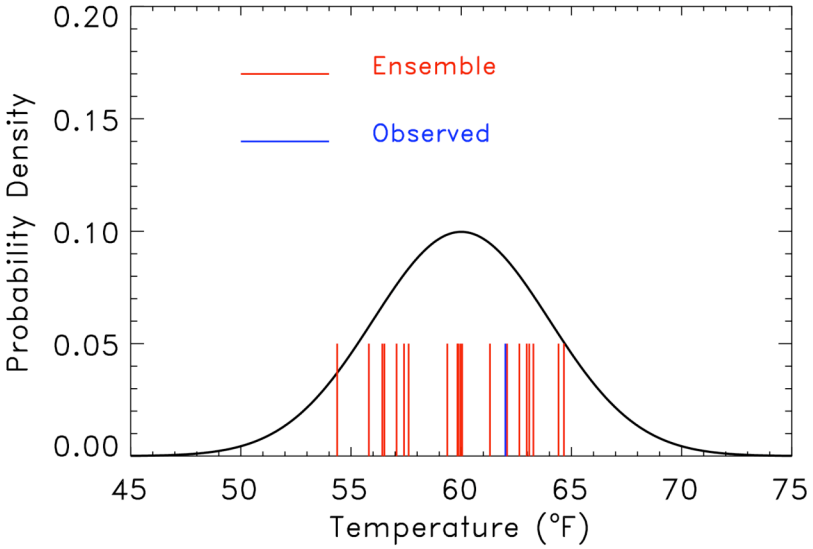


Here, the observed is outside of the range of the ensemble, which was sampled from the pdf shown. Is this a sign of a poor ensemble forecast?

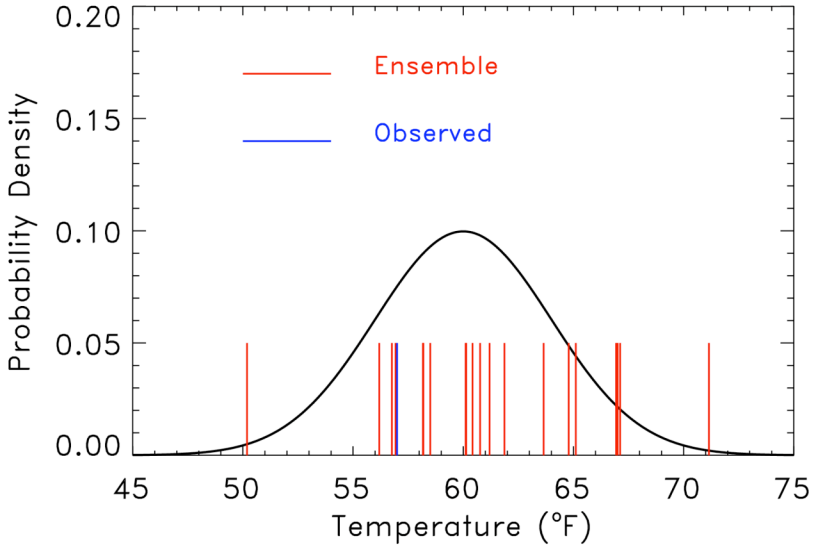
Rank 1 of 21



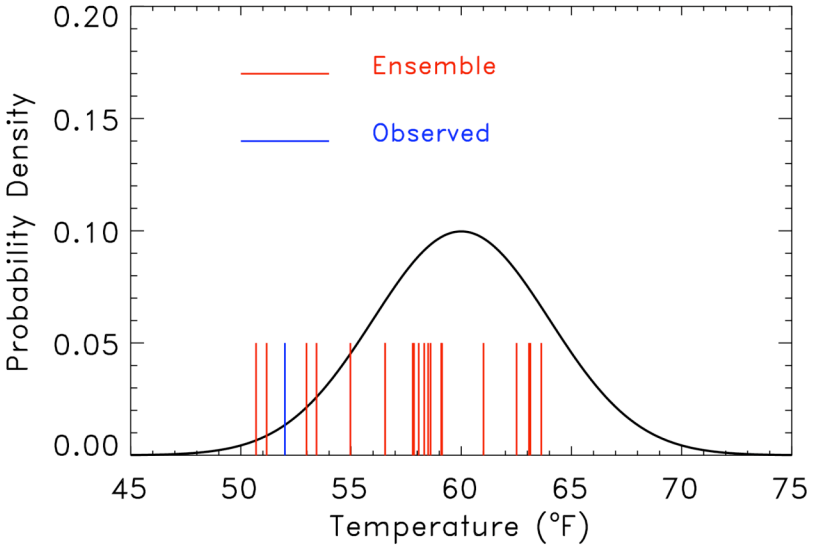
Rank 14 of 21



Rank 5 of 21

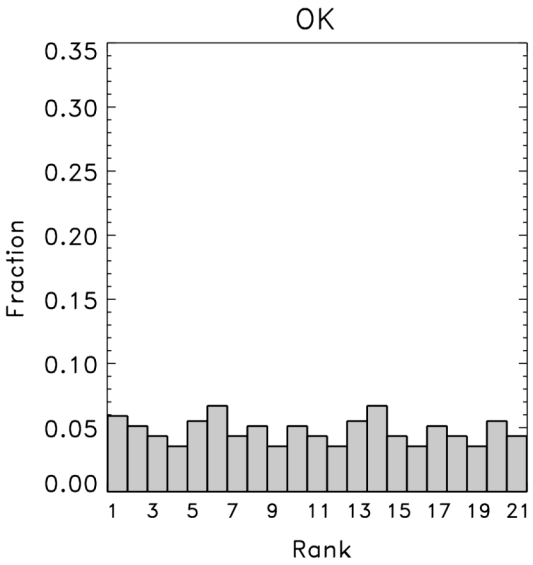


Rank 3 of 21

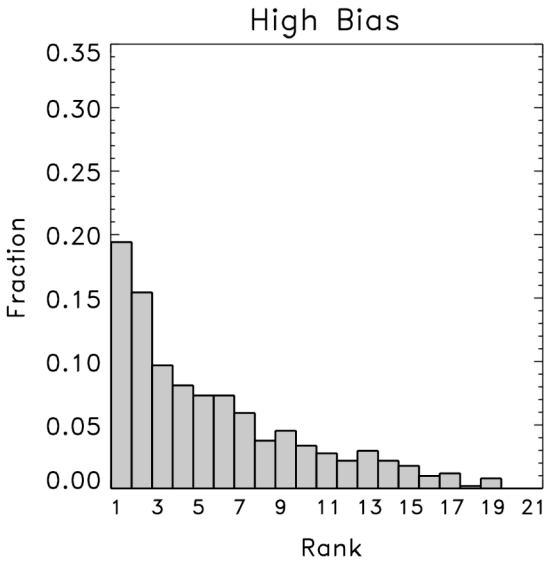


One way of evaluating ensembles: "rank histograms" or "Talagrand diagrams"

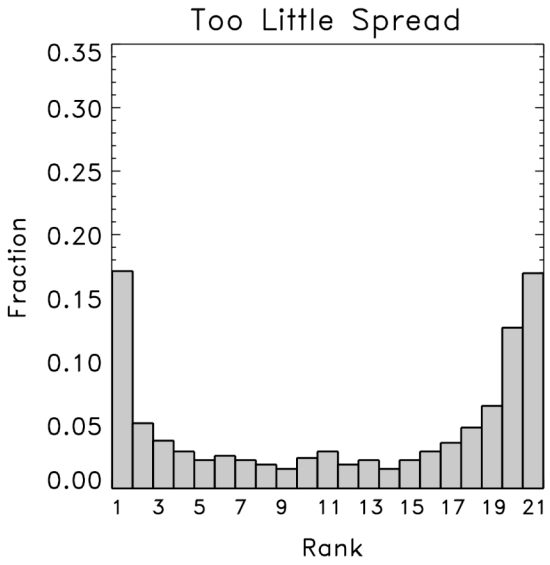
We need lots of samples from many situations to evaluate the characteristics of the ensemble.



Happens when observed is indistinguishable from any other member of the ensemble. Ensemble is "reliable"



Happens when observed too commonly is lower than the ensemble members.

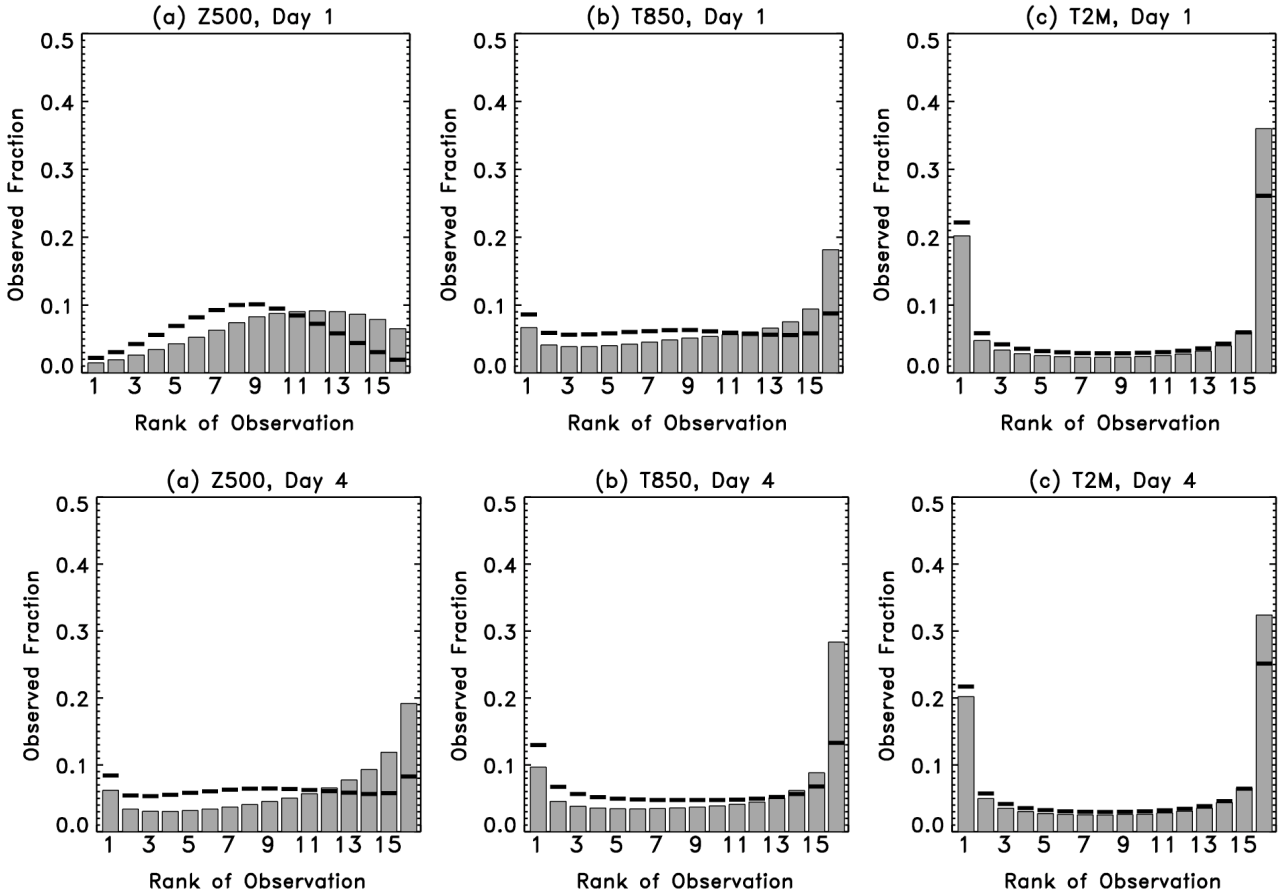


Happens when there are either some low and some high biases, or when the ensemble doesn't spread out enough.

ref: Hamill, *MWR*, March 2001

Rank histograms of Z_{500}, T_{850}, T_{2m}

(from 1998 reforecast version of NCEP GFS)



Solid lines indicate ranks after bias correction. Rank histograms are particularly U-shaped for T_{2M}, which is probably the most relevant of the three plotted here.

Rank histograms for higher dimensions? the "*minimum spanning tree*" histogram

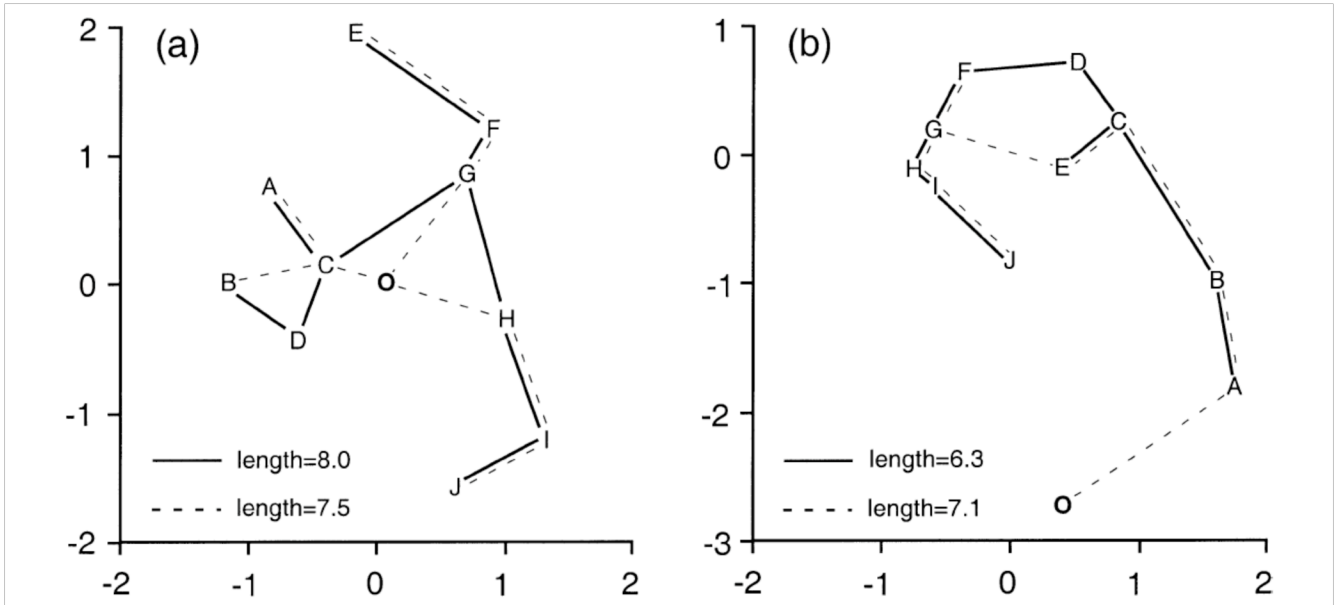


FIG. 1. Hypothetical example MSTs in $K = 2$ dimensions. The $n_{ens} = 10$ ensemble members are labeled A–J, and the corresponding observation is O. Solid lines indicate MSTs for the ensemble as forecast, and dashed lines indicate MSTs that result from the observation being substituted for ensemble member D. (a) A configuration that could result from an overdispersed ensemble, where the observation is interior to the point cloud of the ensemble. (b) A configuration that could result from an underdispersed ensemble and/or a substantial ensemble mean error.

- Solid lines: minimum spanning tree (MST) between 10-member forecasts
- Dashed line: MST when observed O is substituted for member D
- Calculate MST's sum of line segments for all forecasts, and observed replacing each forecast member. Tally rank of pure forecast sum relative to sum where observed replaced a member.
- Repeat for independent samples, build up a histogram

Ref: Wilks, MWR, June 2004. See also Smith and Hansen, MWR, June 2004

Minimum spanning tree histogram interpretation

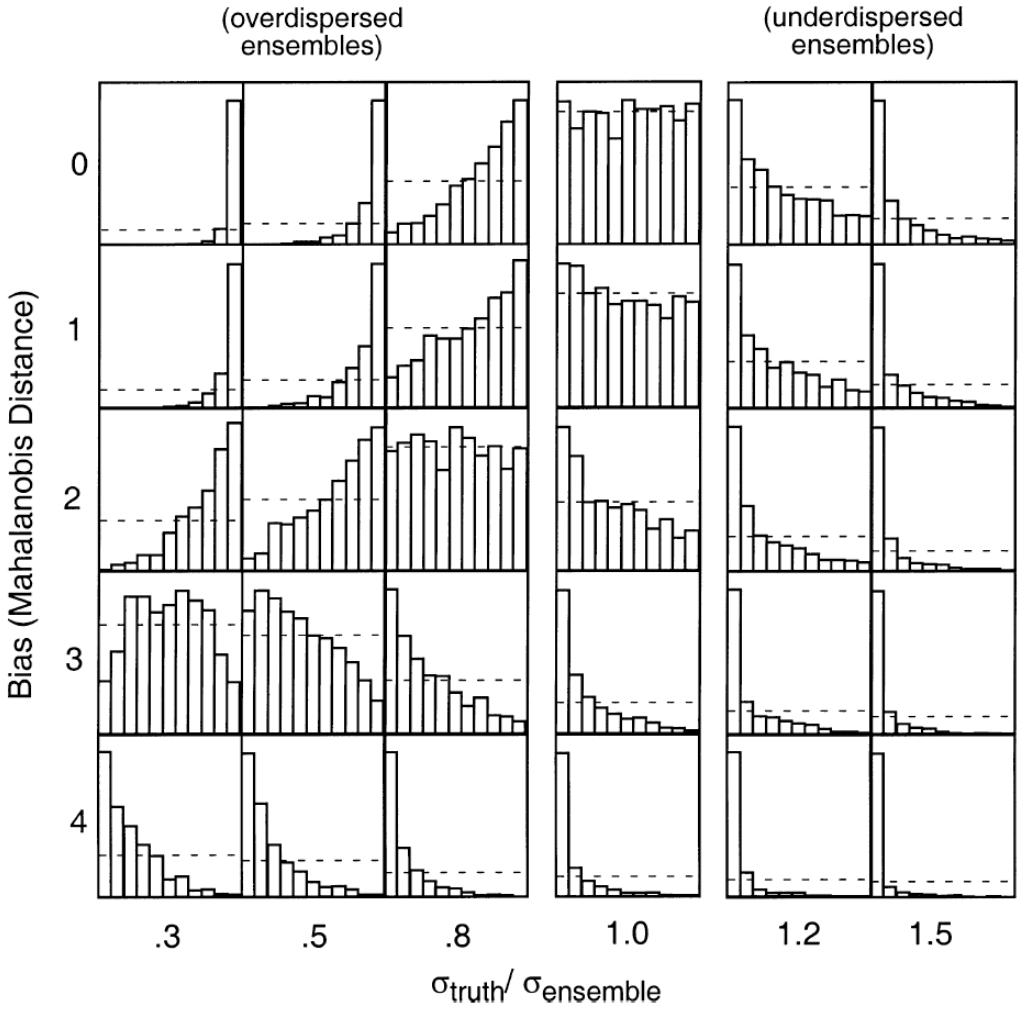
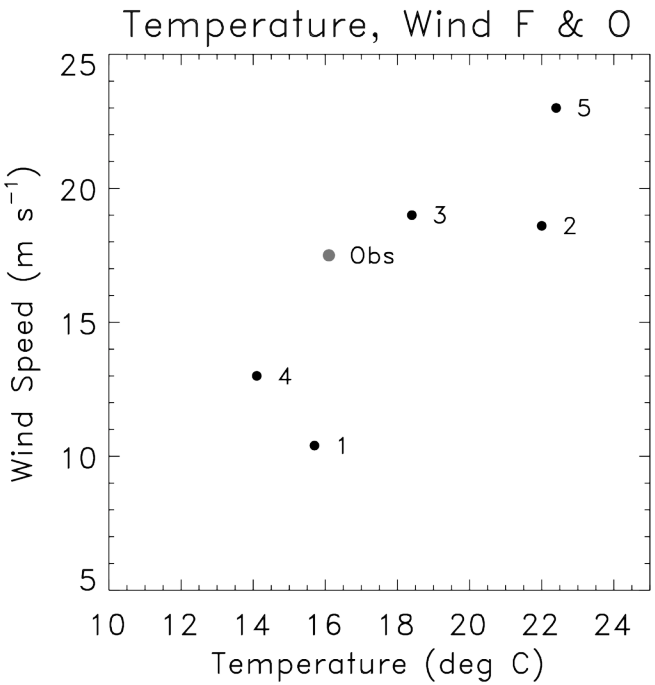


FIG. 2. Behaviors of MST histograms for $n_{ens} = 10$ in $K = 10$ dimensions, as functions of ensemble bias (vertical) and ensemble underdispersion (horizontal), from independent samples of size $n = 1000$. Vertical scales on each histogram have been varied for clarity of presentation, with the level of the expected number per bin under uniformity (1000/11 = 91) indicated in each case by the dashed line.

- Graphical interpretation of MST is different than it is for uni-dimensional rank histogram, a disadvantage.
- Is there a multi-dimensional rank histogram with the same geographic interpretation as the scalar rank histogram?

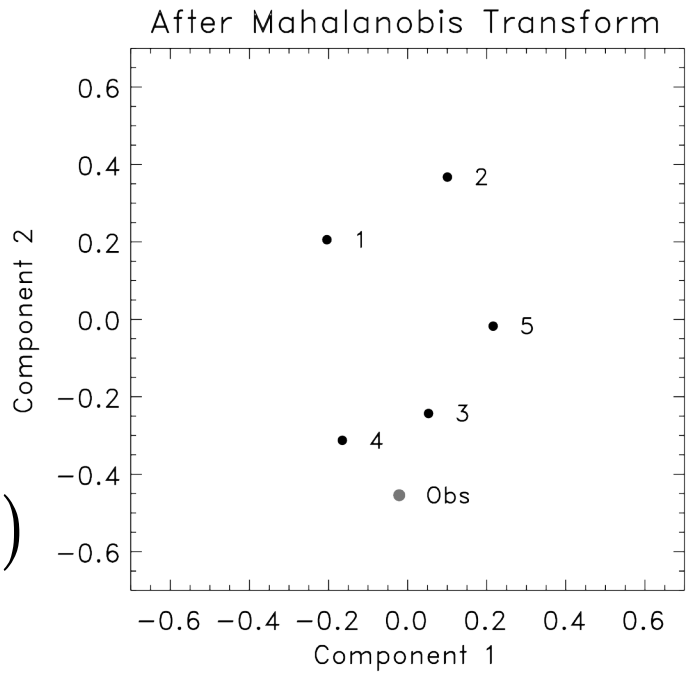
Ref: Wilks, MWR, June 2004. See also Smith and Hansen, MWR, June 2004

Multi-variate rank histogram



"Mahalanobis" transform (S is forecasts' sample covariance)

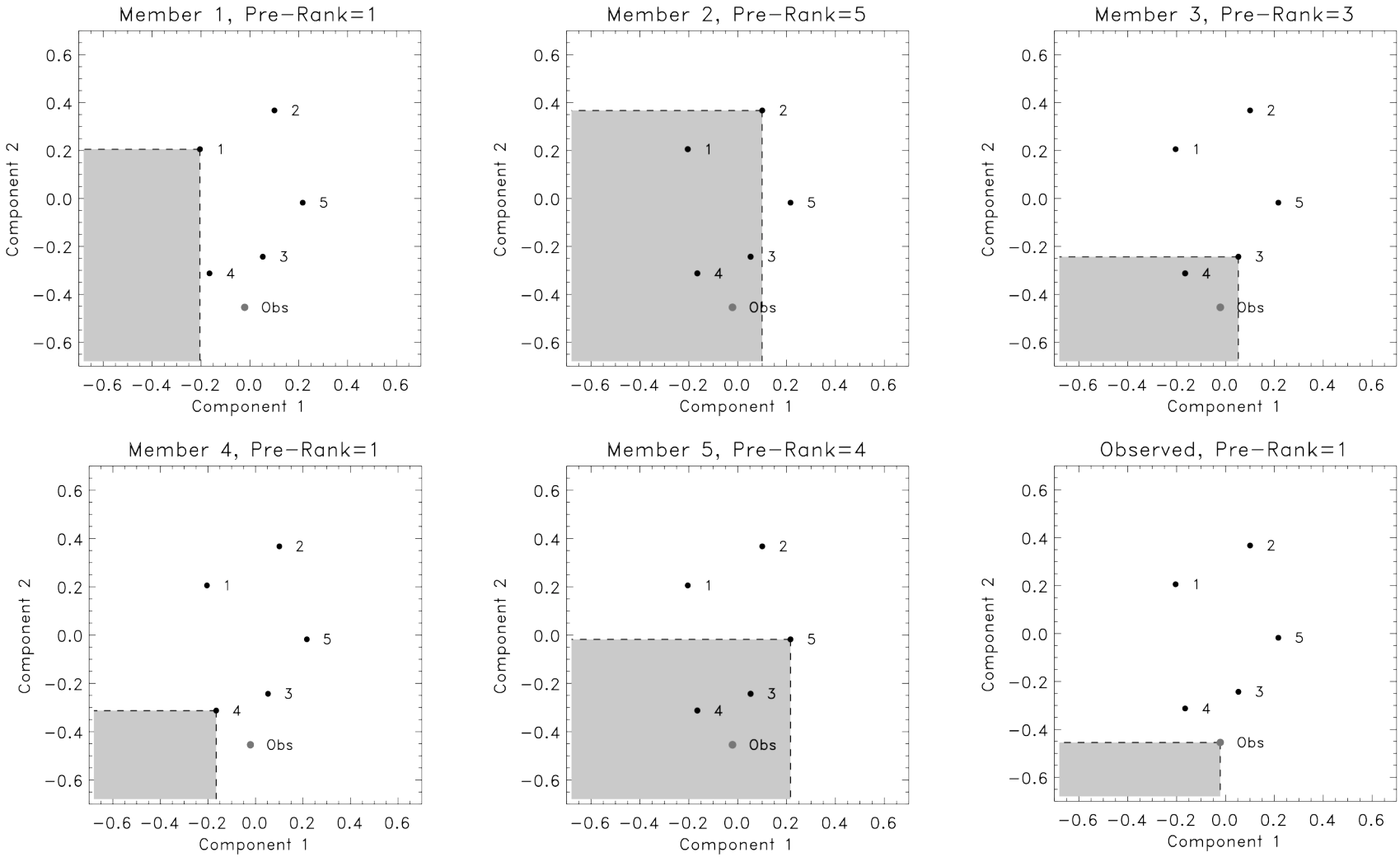
$$z_i = [S]^{-1/2}\left(x_i - \bar{x}\right)$$



- Standardize and rotate using Mahalanobis transformation (see Wilks 2006 text).
- For each of *n* members of forecast and observed, define "pre-rank" as the number of vectors to its lower left (a number between 1 and *n*+1)
- The multi-variate rank is the rank of the observation pre-rank, with ties resolved at random
- Composite multi-variate ranks over many independent samples and plot rank histogram.
- Same interpretation as scalar rank histogram (e.g., U-shape = under-dispersive).

based on Tilmann Gneiting's presentation at Probability and Statistics, 2008 AMS Annual Conf., New Orleans.

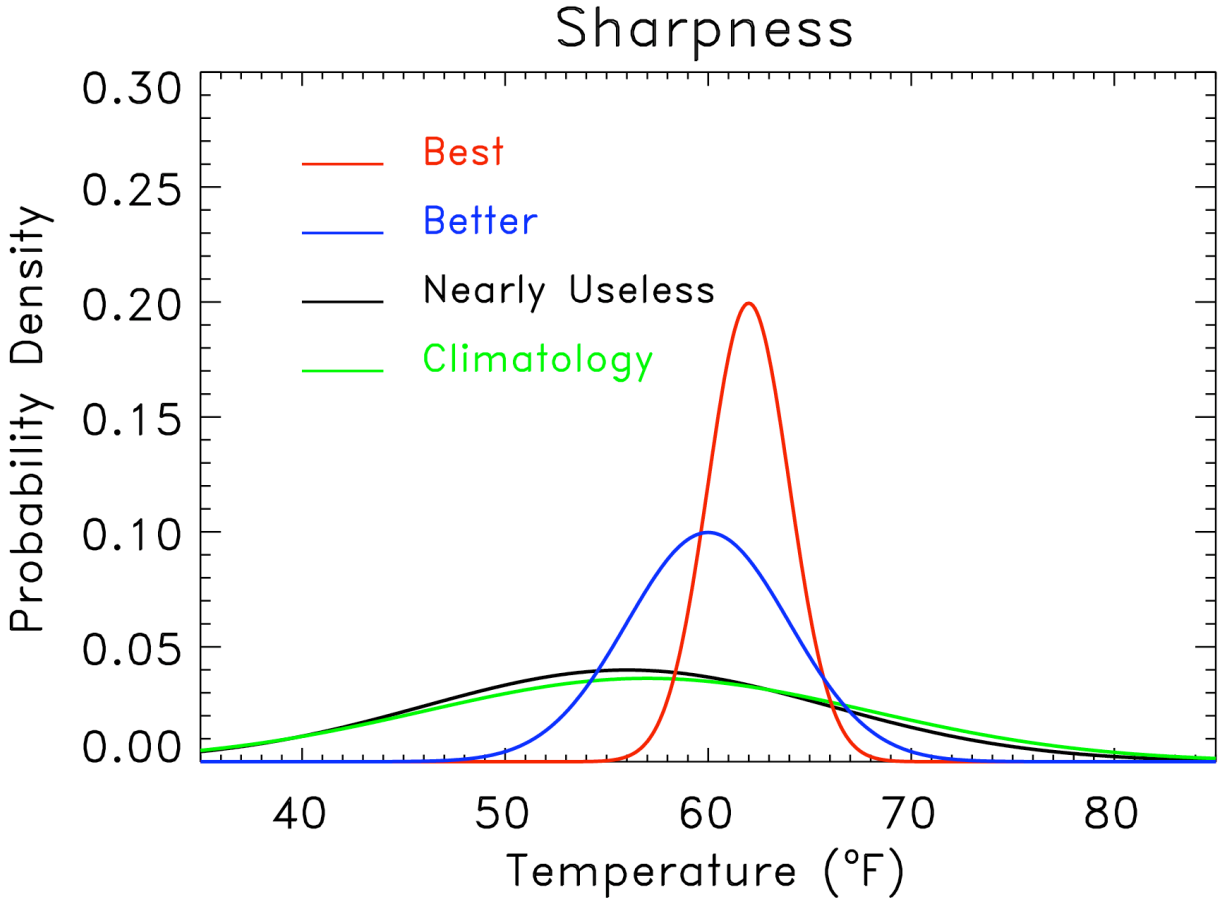
Multi-variate rank histogram calculation



F_1, F_2, F_3, F_4, F_5, O pre-ranks: [1, 5, 3, 1, 4, 1] → sorted: obs = either rank 1, 2, or 3 with p=1/3.

based on Tilmann Gneiting's presentation at Probability and Statistics, 2008 AMS Annual Conf., New Orleans

Rank histograms tell us about reliability - but what else is important?

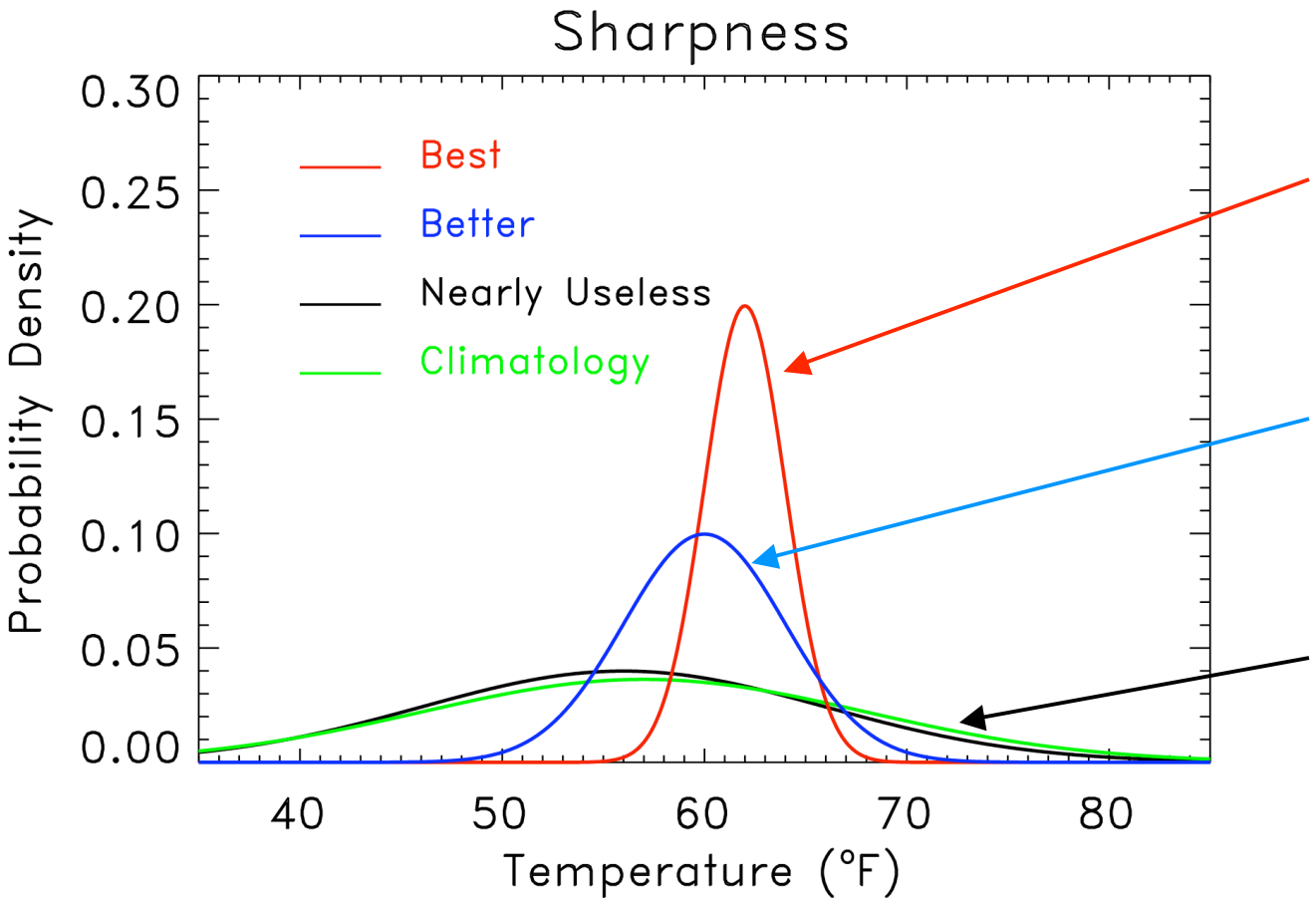


"Sharpness" measures the specificity of the probabilistic forecast. Given two reliable forecast systems, the one producing the sharper forecasts is preferable.

But: don't want sharp if not reliable. Implies unrealistic confidence.

"Spread-skill" relationships are important, too.

Small-spread ensemble forecasts should have less ensemble-mean error than large-spread forecasts.

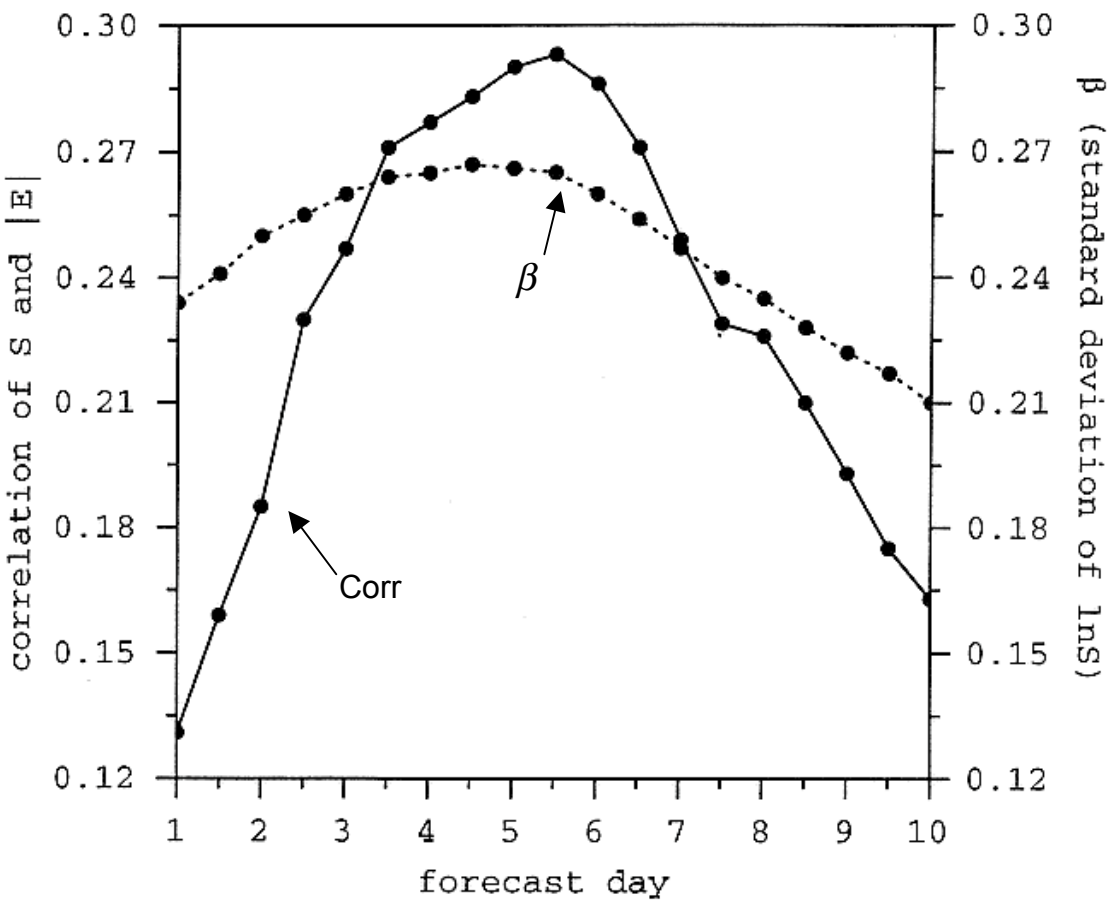


ensemble-mean error from a sample of this pdf on avg. should be low.

ensemble-mean error should be moderate on avg.

ensemble-mean error should be large on avg.

Spread-skill for 1990's NCEP GFS



At a given grid point, spread S is assumed to be a random variable with a lognormal distribution

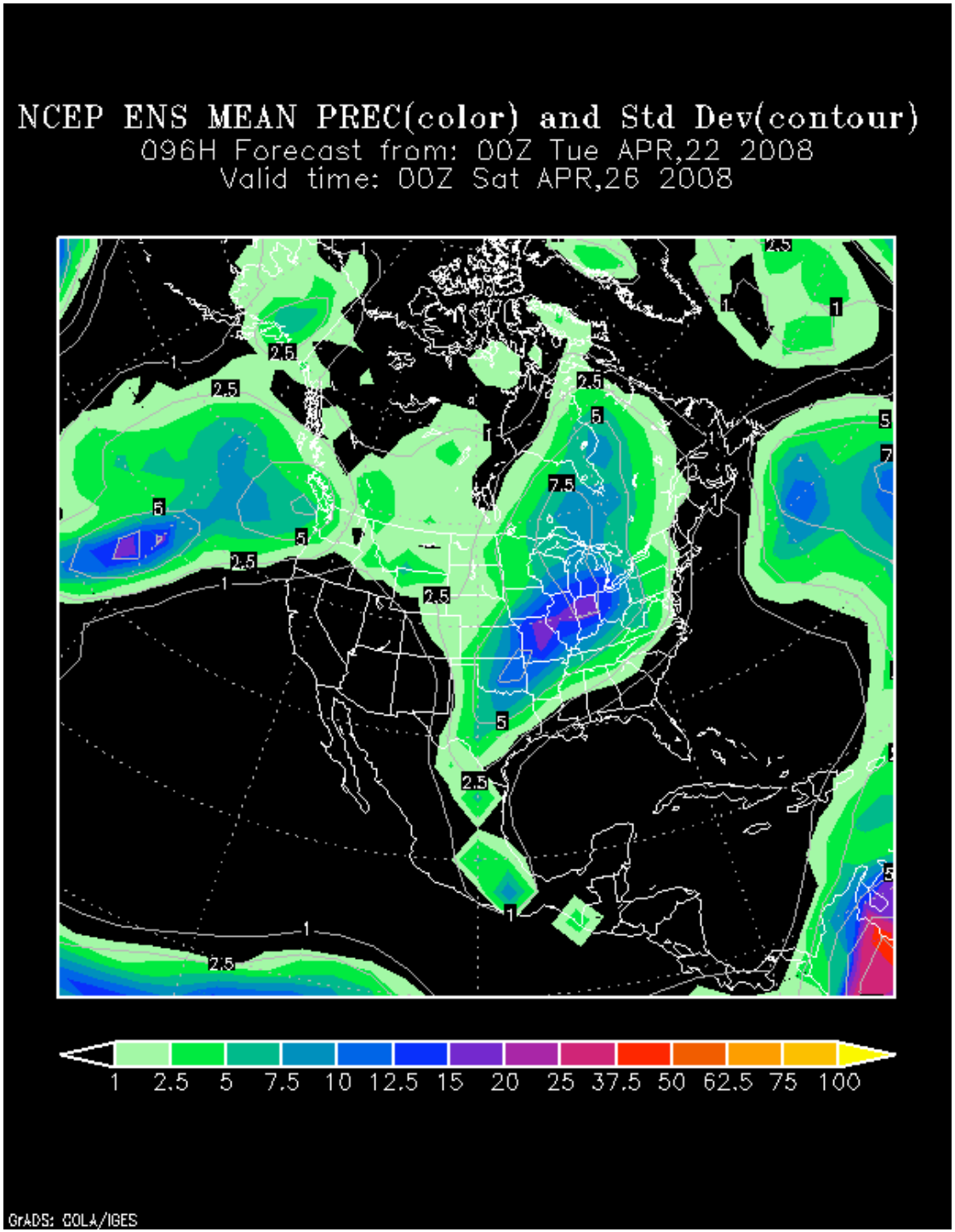
$$\ln S \sim N\left(\ln S_m, \beta\right)$$

where S_m is the mean spread and β is its standard deviation.

As β increases, there is a wider range of spreads in the sample. One would expect then the possibility for a larger spread-skill correlation.

Here β and spread-skill correlation are shown for late 1990's NCEP global forecast model.

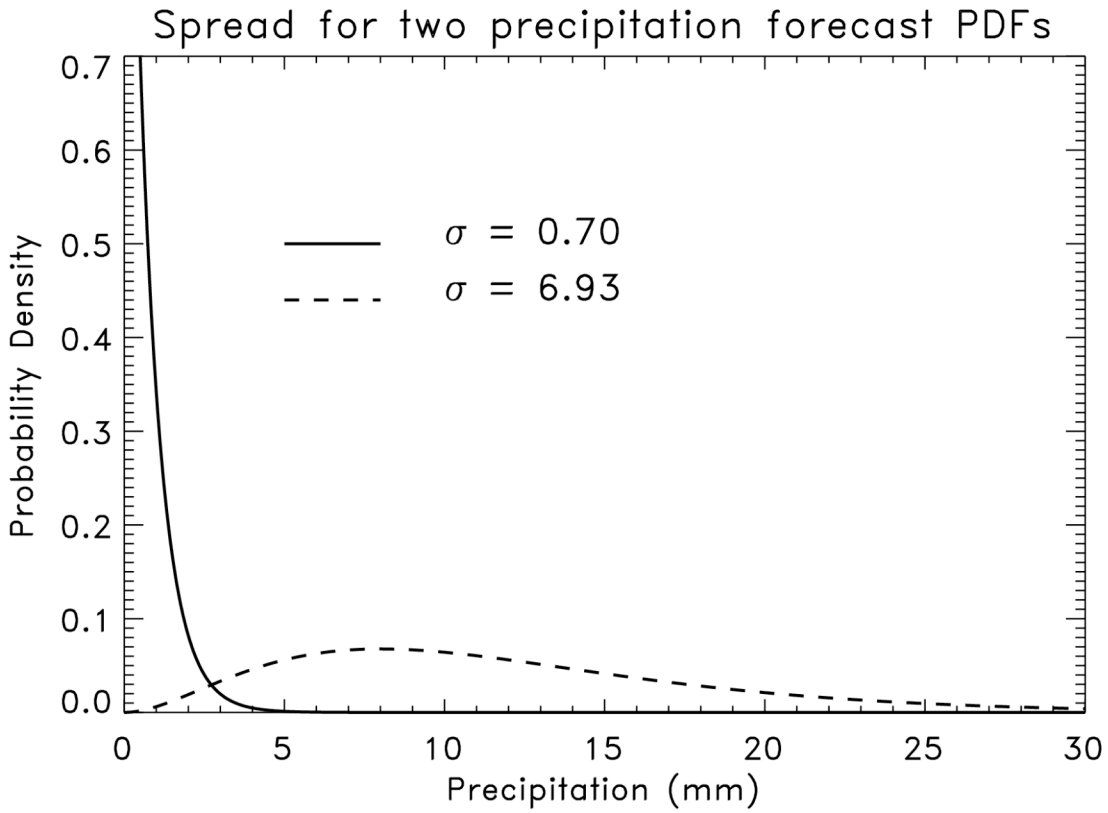
from Whitaker and Loughe, *MWR*, Dec. 1998



Ensemble mean and standard deviation of precipitation

- Mean colored, standard deviation in contours. Notice the strong similarity.

Spread-skill and precipitation forecasts



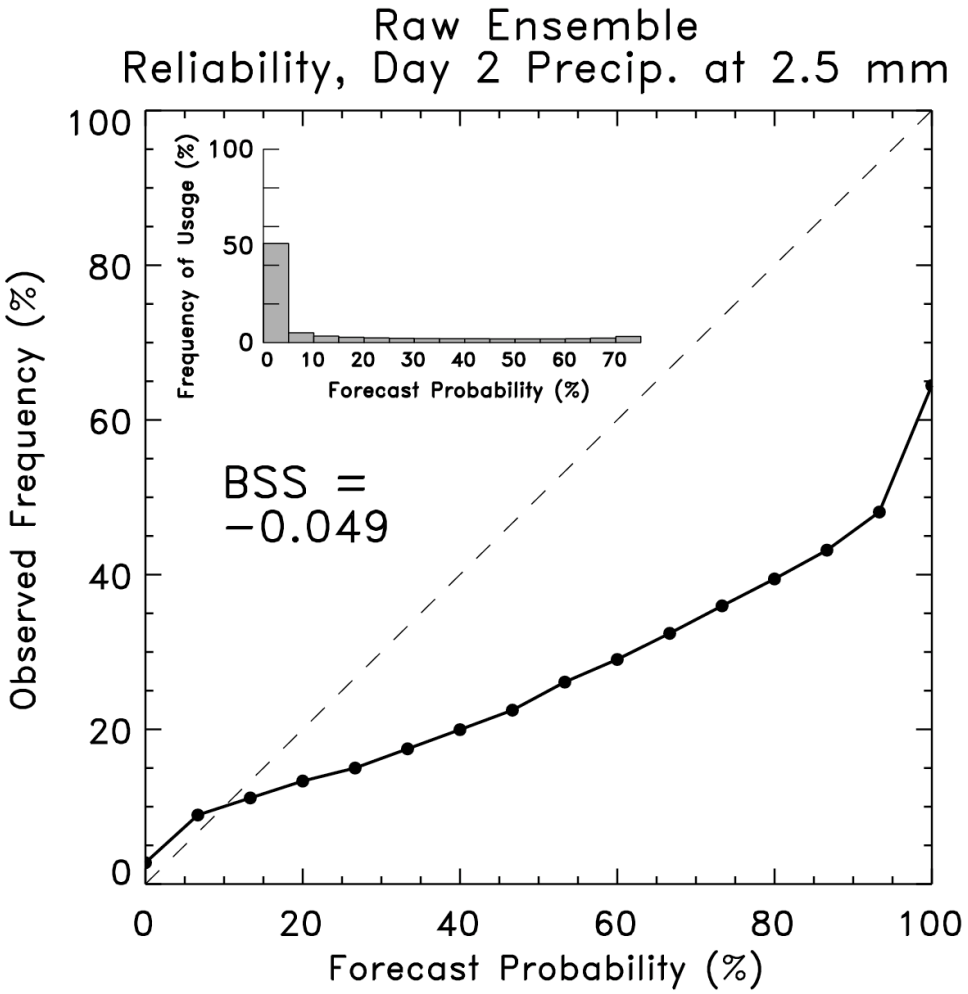
True spread-skill relationships harder to diagnose if forecast PDF is non-normally distributed, as they are typically for precipitation forecasts.

Commonly, spread is no longer independent of the mean value; it's larger when the amount is larger.

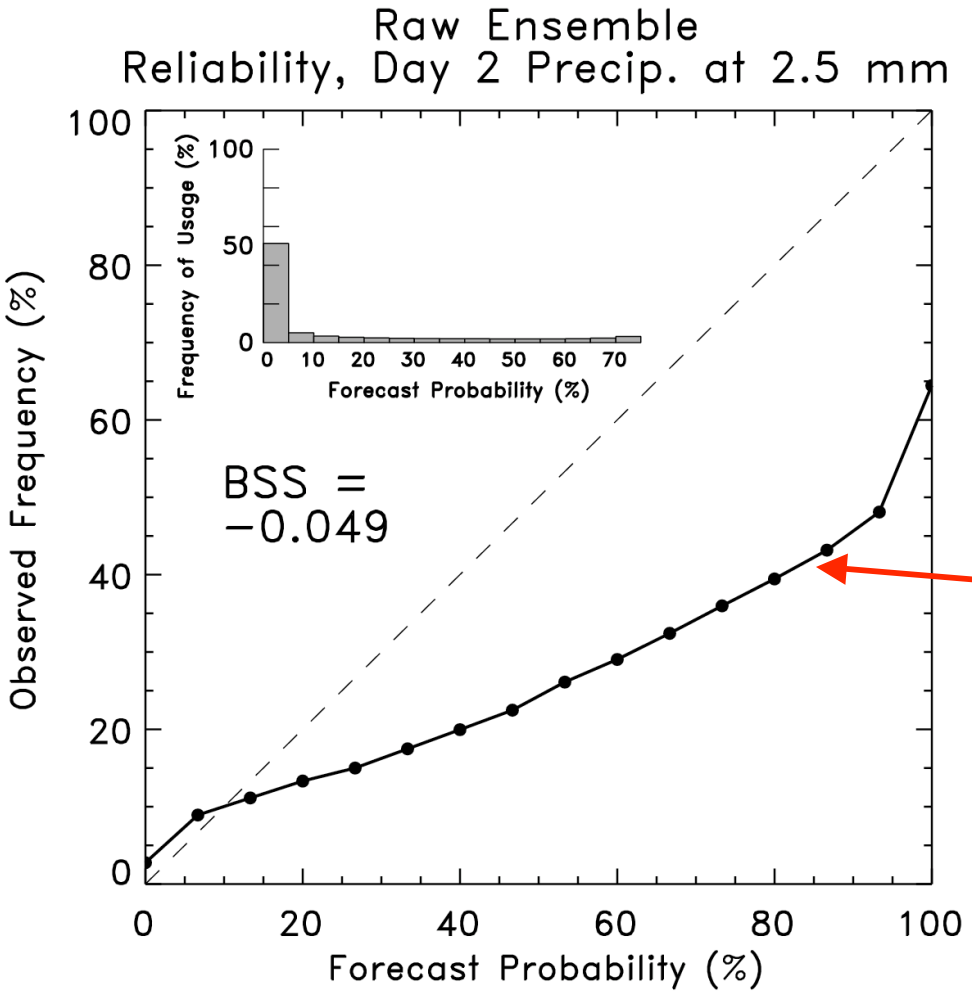
Hence, you get an apparent spread-skill relationship, but this may reflect variations in the mean forecast rather than real spread-skill.

See Hamill and Colucci, *MWR*, 1998 for more discussion on this

Reliability diagrams



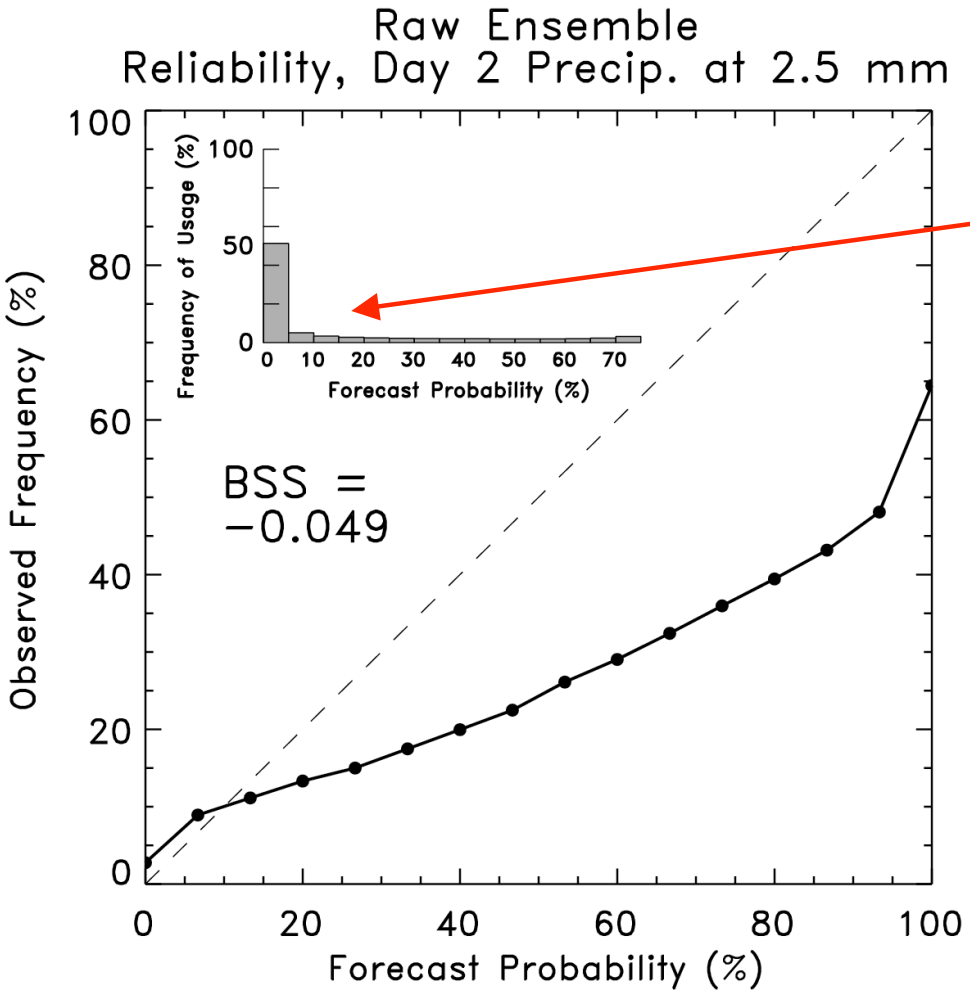
Reliability diagrams



Curve tells you what the observed frequency was each time you forecast a given probability. This curve ought to lie along $y = x$ line. Here this shows the ensemble-forecast system over-forecasts the probability of light rain.

Ref: Wilks text, *Statistical Methods in the Atmospheric Sciences*

Reliability diagrams

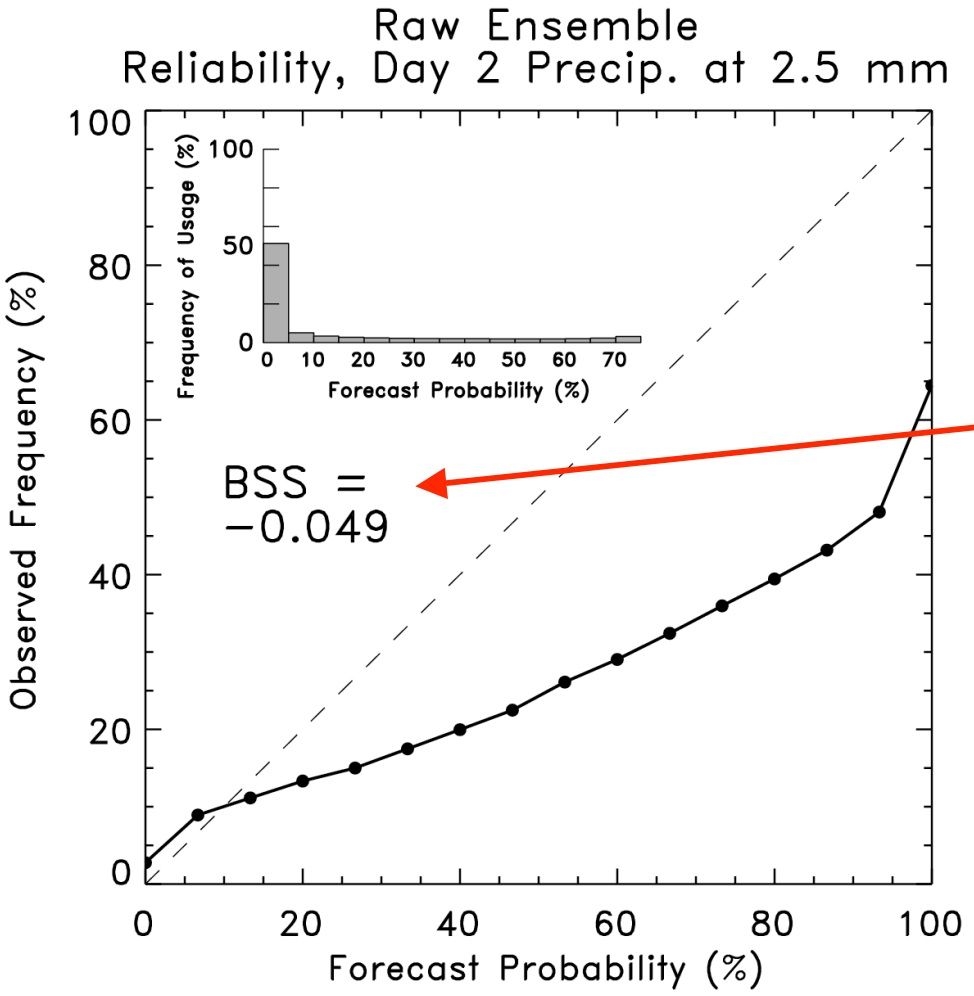


Inset histogram tells you how frequently each probability was issued.

Perfectly sharp: frequency of usage populates only 0% and 100%.

Ref: Wilks text, *Statistical Methods in the Atmospheric Sciences*

Reliability diagrams



BSS = Brier Skill Score

$$BSS = \frac{BS(CLimo) - BS(Forecast)}{BS(CLimo) - BS(Perfect)}$$

BS(•) measures the Brier Score, which you can think of as the squared error of a probabilistic forecast.

Perfect: *BSS* = 1.0
Climatology: *BSS* = 0.0

Ref: Wilks text, *Statistical Methods in the Atmospheric Sciences*

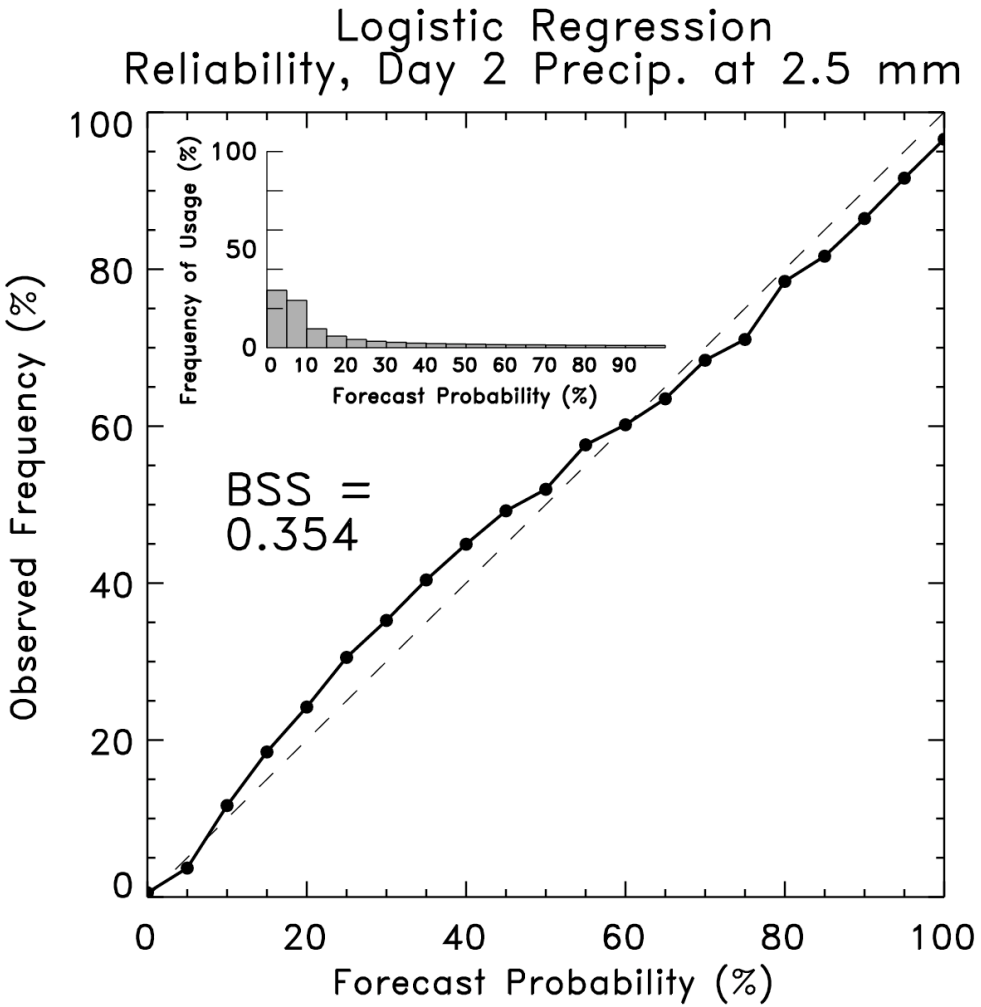
Brier score

- Define an event, e.g., obs. precip > 2.5 mm.
- Let P_i^f be the forecast probability for the *i*th forecast case.
- Let O_i be the observed probability (1 or 0). Then

$$BS(\text{forecast}) = \frac{1}{ncases} \sum_{i=1}^{ncases} \left(P_i^f - O_i \right)^2$$

(So the Brier score is the averaged squared error of the probabilistic forecast)

Reliability after post-processing

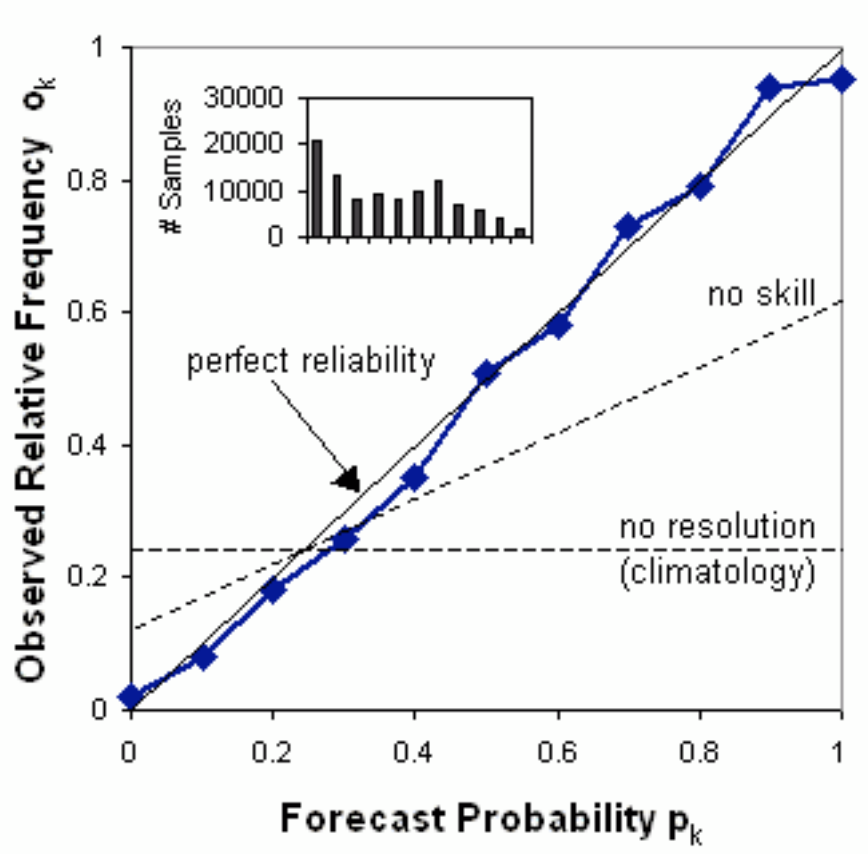


Statistical correction of forecasts using a long, stable set of prior forecasts from the same model (like in MOS). More on this in reforecast seminar.

Ref: Hamill et al., MWR, Nov 2006

"Attributes diagram"

(a slight variant of the reliability diagram)



$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

reliability *resolution* *uncertainty*

$$\text{BSS} = \frac{\text{"Resolution"} - \text{"Reliability"}}{\text{"Uncertainty"}}$$

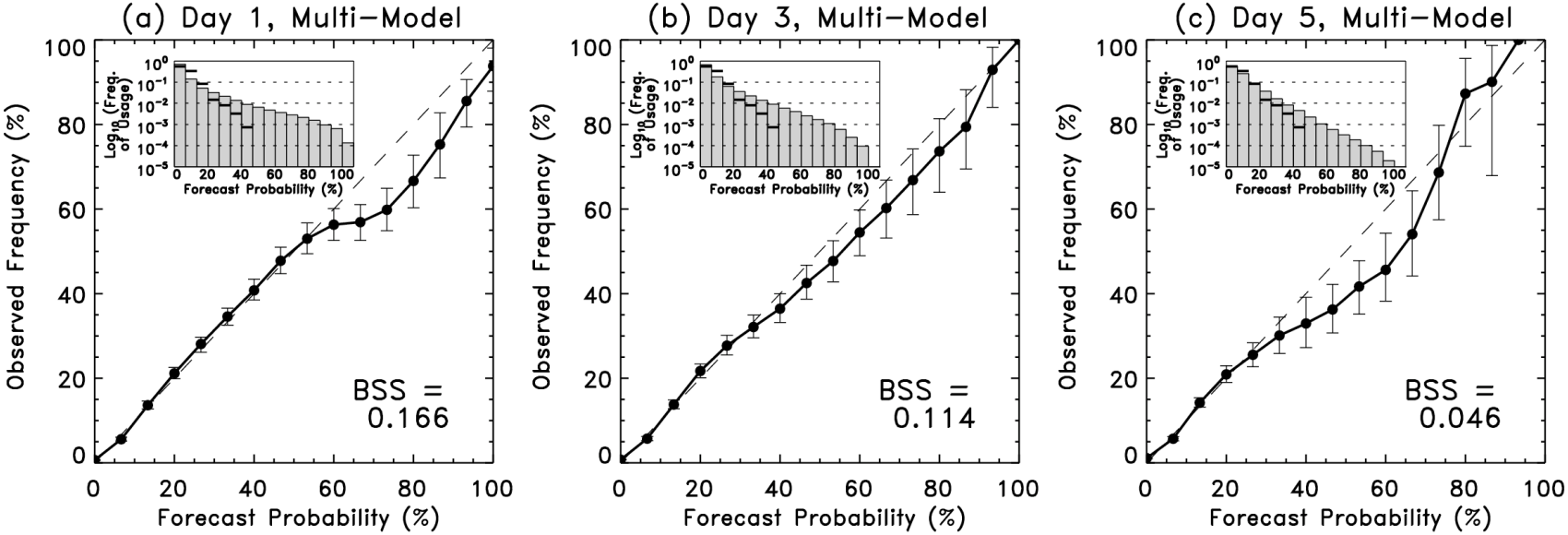
Uncertainty term always positive, so probability forecasts will exhibit positive skill if resolution term is larger in absolute value than reliability term. Geometrically, this corresponds to points on the attributes diagram being closer to 1:1 perfect reliability line than horizontal no-resolution line (from Wilks text, 2006, chapter 7)

Note, however, that this geometric interpretation of the attributes diagram is correct only if all samples used to populate the diagram are drawn from the same climatological distribution. If one is mixing samples from locations with different climatologies, this interpretation is no longer correct! (for more on what underlies this issue, see Hamill and Juras, Oct 2006 *QJRMS*)

www.bom.gov.au/bmrc/wefor/staff/eee/verif/ReliabilityDiagram.gif, from Beth Ebert's verification web page, http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html based on Hsu and Murphy, 1986, *Int'l Journal of Forecasting*

Proposed modifications to reliability diagrams

12-h accumulated forecasts, 5-mm threshold, over US



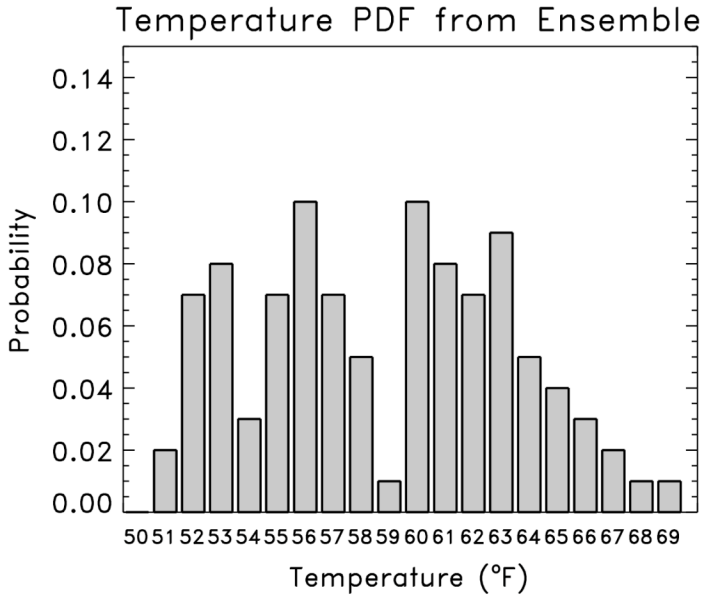
- Block-bootstrap techniques (each forecast day is a block) to provide confidence intervals. See also Hamill, *WAF*, April 1999, and Bröcker and Smith, *WAF*, June 2007.
- Distribution of climatological forecasts plotted as horizontal bars on the inset histogram. Helps explain why there is small skill for a forecast that appears so reliable (figure from Hamill et al., *MWR*, 2008 to appear).

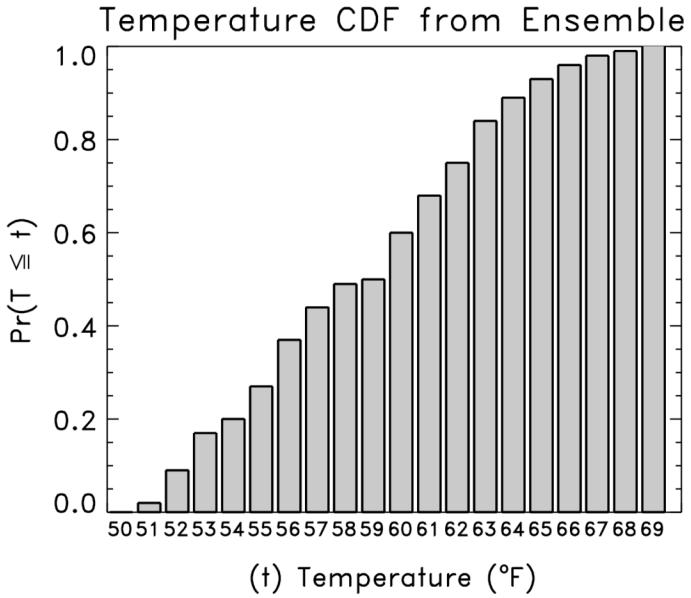
Continuous ranked probability score

Start with cumulative distribution function (CDF)

$F^f(x) = \Pr\{X \leq x\}$

where X is the random variable, x is some specified threshold.

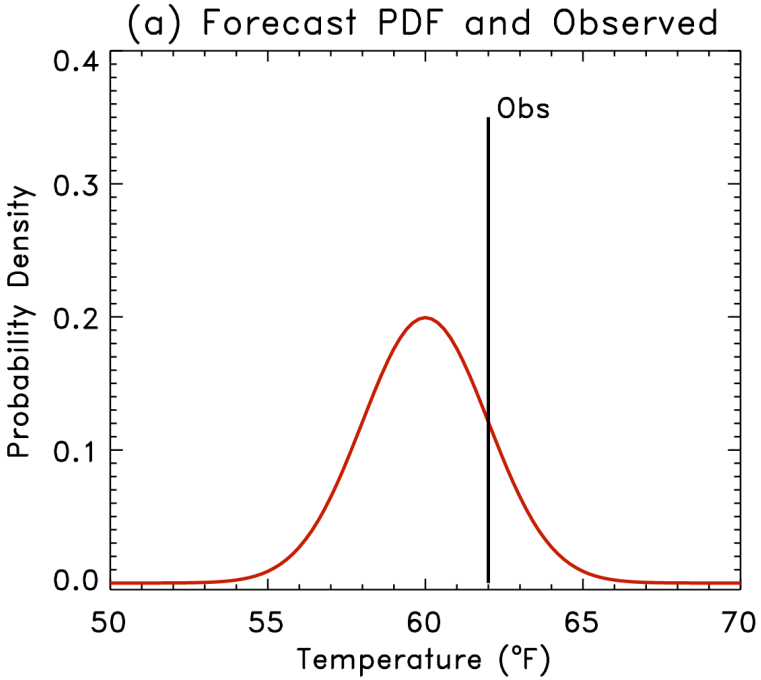


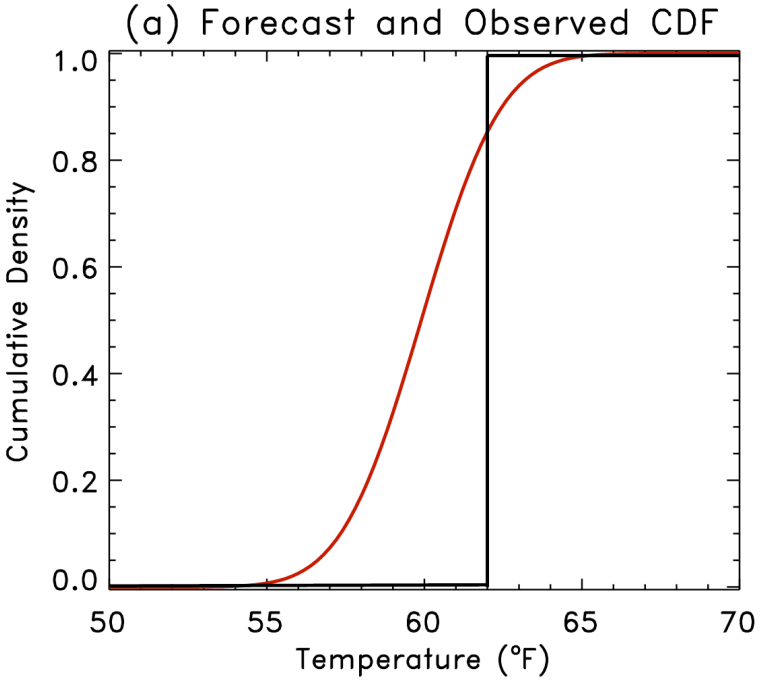


Continuous ranked probability score

- Let $F_i^f(x)$ be the forecast probability CDF for the *i*th forecast case.
- Let $F_i^o(x)$ be the observed probability CDF (Heaviside function).

$$CRPS(forecast) = \frac{1}{ncases} \sum_{i=1}^{ncases} \int_{x=-\infty}^{x=-\infty} \left(F_i^f(x) - F_i^o(x) \right)^2 dx$$



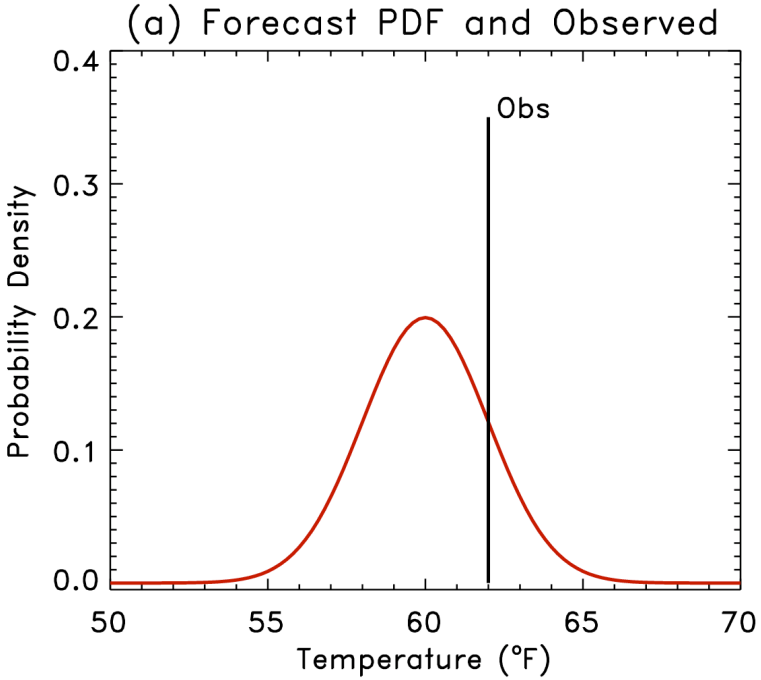


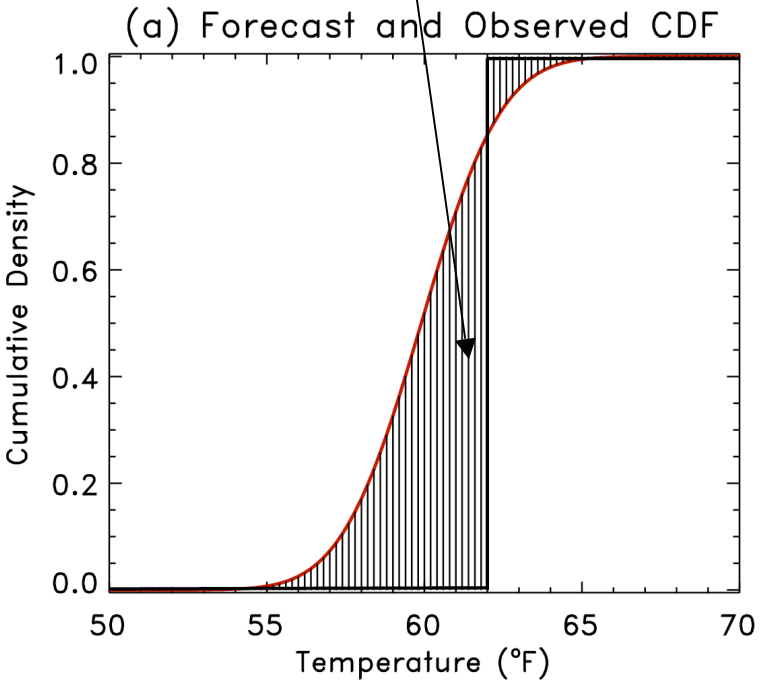
Continuous ranked probability score

- Let $F_i^f(x)$ be the forecast probability CDF for the *i*th forecast case.
- Let $F_i^o(x)$ be the observed probability CDF (Heaviside function).

$$CRPS(forecast) = \frac{1}{ncases} \sum_{i=1}^{ncases} \int_{x=-\infty}^{x=-\infty} \left(F_i^f(x) - F_i^o(x) \right)^2 dx$$

(squared)





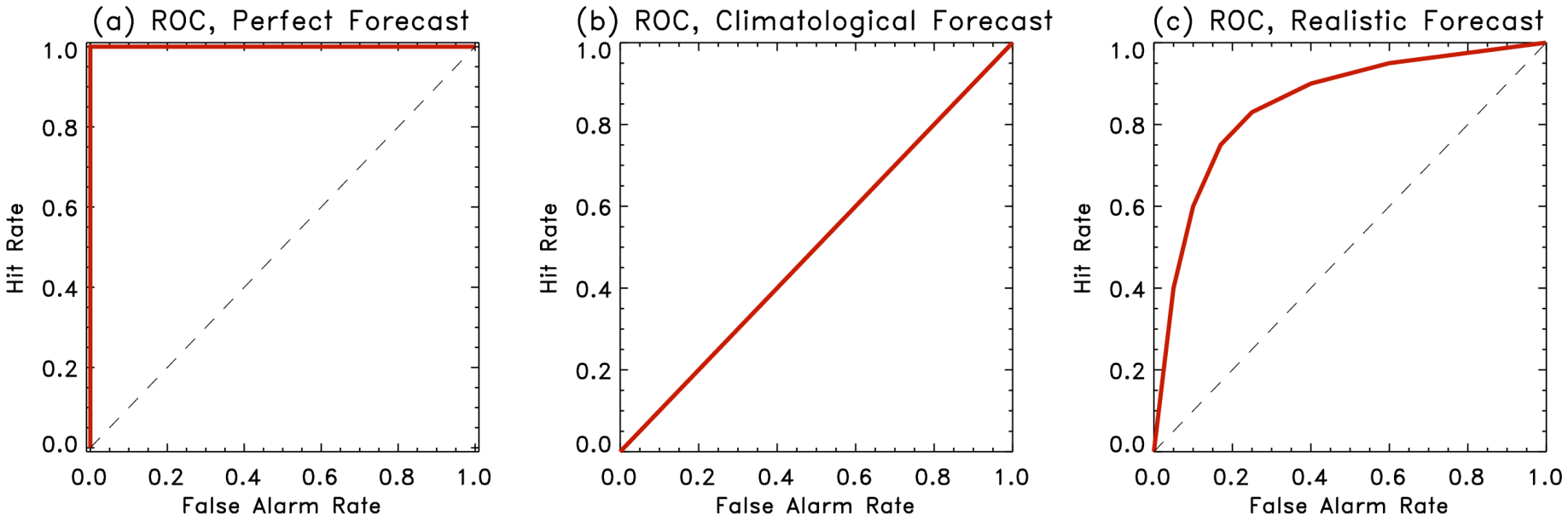
Continuous ranked probability *skill* score (CRPSS)

Like the Brier score, it's common to convert this to a skill score by normalizing by the skill of climatology, or some other reference.

$$CRPSS = \frac{\overline{CRPS}(forecast) - \overline{CRPS}(climo)}{\overline{CRPS}(perfect) - \overline{CRPS}(climo)}$$

Ref: Wilks 2006 text

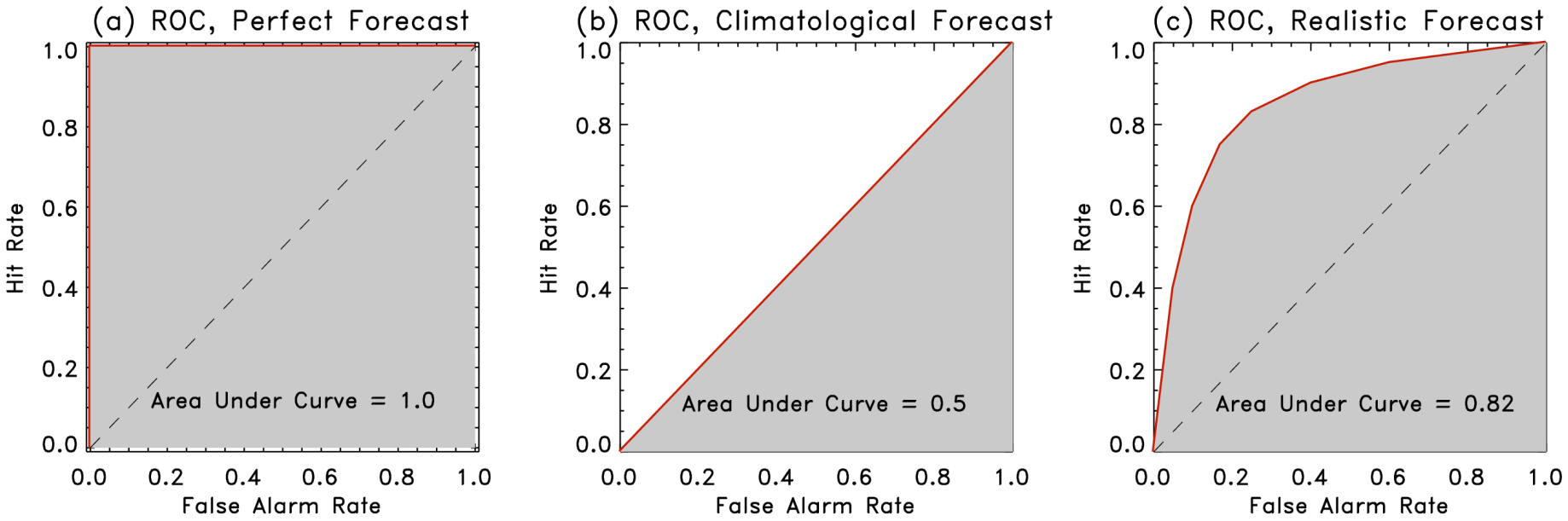
Relative operating characteristic (ROC)



Measures tradeoff of Type I statistical errors (incorrect rejection of null hypothesis) against Type II (incorrect acceptance of alternative) as decision threshold is changed.

see Mason, 1982, *Austr. Meteor. Mag*, and Harvey et al., 1992 *MWR* for a review

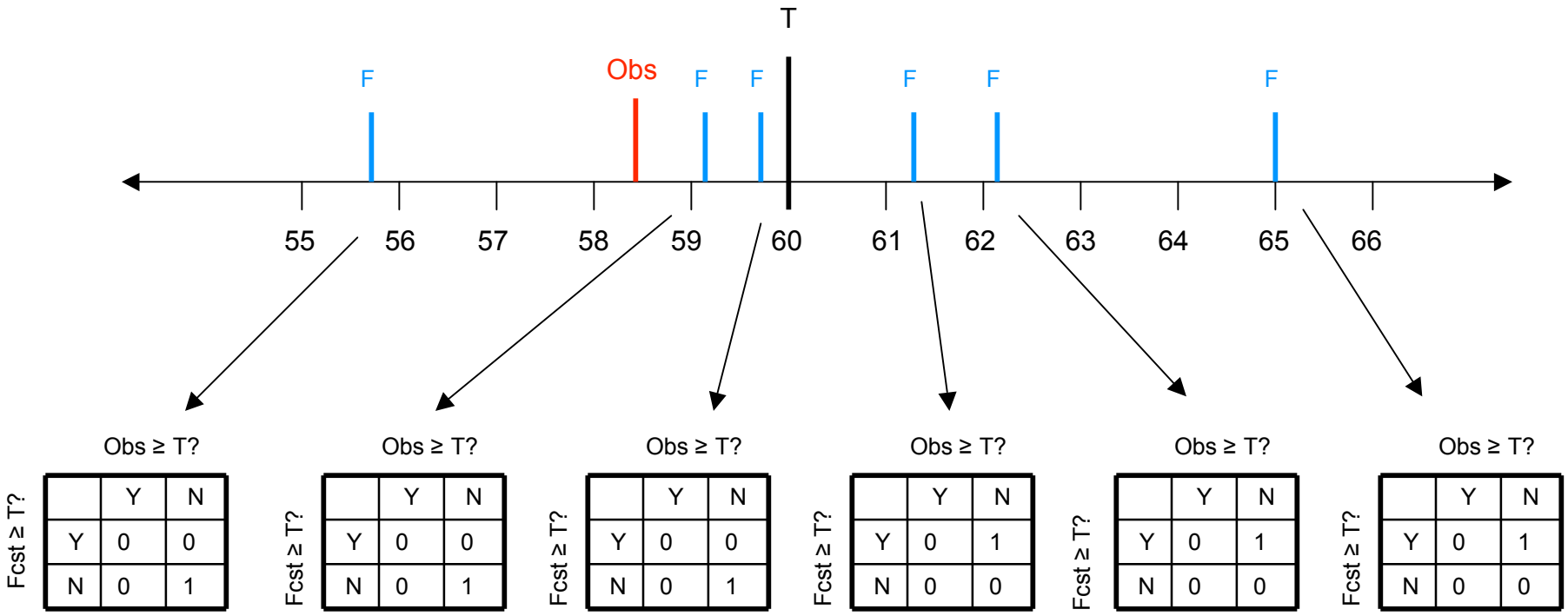
Relative operating characteristic (ROC)



$$ROCSS = \frac{AUC_f - AUC_{clim}}{AUC_{perf} - AUC_{clim}} = \frac{AUC_f - 0.5}{1.0 - 0.5} = 2AUC_f - 1$$

Method of calculation of ROC: parts 1 and 2

(1) Build contingency tables for each sorted ensemble member



(2) Repeat the process for other locations, dates, building up contingency tables for sorted members.

Method of calculation of ROC: part 3

(3) Get hit rate and false alarm rate for each from contingency table for each sorted ensemble member.

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	H	F
N	M	C

$$HR = H / (H+M) \qquad FAR = F / (F+C)$$

Sorted Member 1

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	1106	3
N	5651	73270

HR = 0.163
FAR = 0.000

Sorted Member 2

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	3097	176
N	3630	73097

HR = 0.504
FAR = 0.002

Sorted Member 3

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	4020	561
N	2707	72712

HR = 0.597
FAR = 0.007

Sorted Member 4

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	4692	1270
N	2035	72003

HR = 0.697
FAR = 0.017

Sorted Member 5

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	5297	2655
N	1430	70618

HR = 0.787
FAR = 0.036

Sorted Member 6

Fcst ≥ T? \ Obs ≥ T?	Y	N
Y	6603	44895
N	124	28378

HR = 0.981
FAR = 0.612

Method of calculation of ROC: parts 3 and 4

HR = 0.163
FAR = 0.000

HR = 0.504
FAR = 0.002

HR = 0.597
FAR = 0.007

HR = 0.697
FAR = 0.017

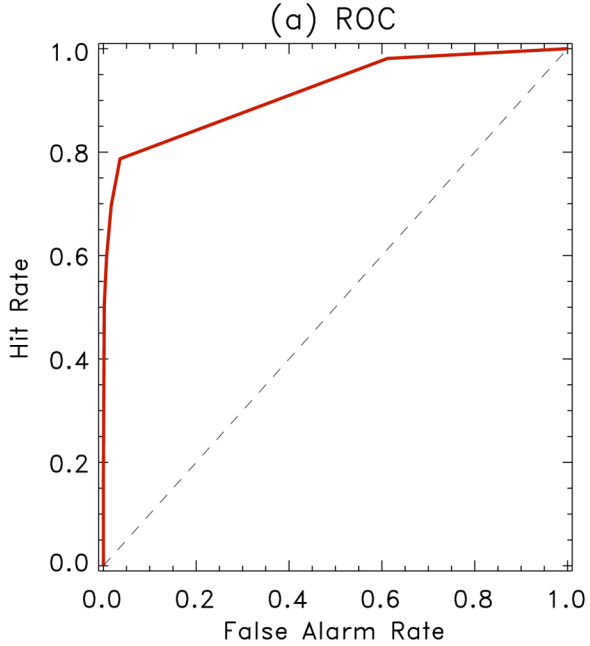
HR = 0.787
FAR = 0.036

HR = 0.981
FAR = 0.612

HR = [0.000, 0.163, 0.504, 0.597, 0.697, 0.787, 0.981, 1.000]

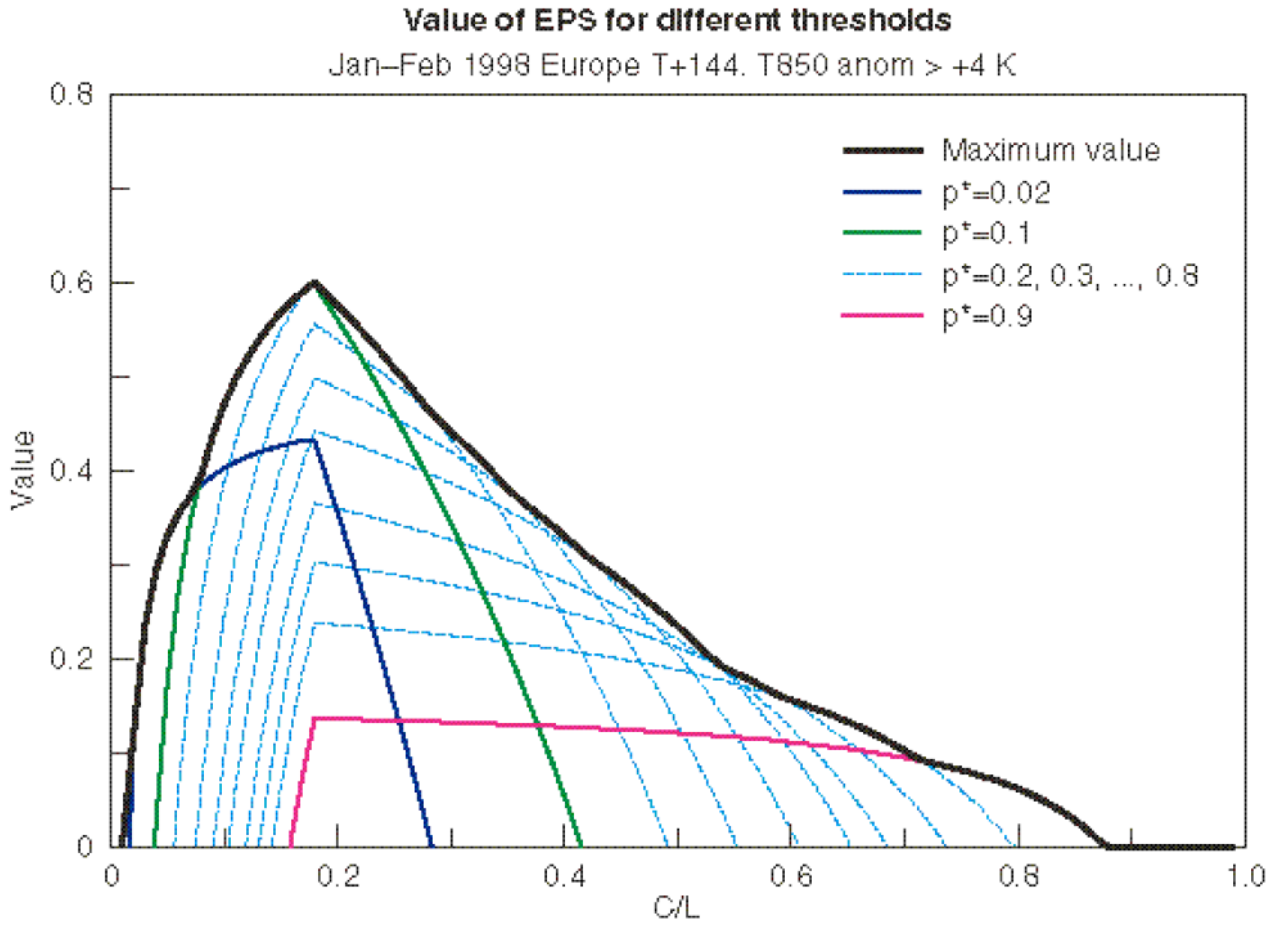
FAR = [0.000, 0.000, 0.002, 0.007, 0.017, 0.036, 0.612, 1.000]

(4) Plot hit rate vs. false alarm rate



Potential economic value diagrams

Motivated by search for a metric that relates ensemble forecast performance to things that customers will actually care about.



These diagrams tell you the potential economic value of your ensemble forecast system applied to a particular forecast aspect. Perfect forecast has value of 1.0, climatology has value of 1.0. Value differs with user's cost/loss ratio.

from Zhu et al. review article, *BAMS*, 2001

Potential economic value: calculation method

Contingency table indicating the costs and losses accrued by the use of weather forecasts, depending on forecast and observed events.

Observation \ Forecast/action	Yes	No	
Yes	Hit (h) Mitigated loss ($C + L_u$)	Miss (m) Loss ($L = L_p + L_u$)	$h + m = \bar{o}$
No	False Alarm (f) Cost (C)	Correct rejection (c) No cost (N)	$f + c = 1 - \bar{o}$

Assumes decision maker alters actions based on weather forecast info.

C = Cost of protection
$L = L_p + L_u$ = total cost of a loss, where …
L_p = Loss that can be protected against
L_u = Loss that can't be protected against.
N = No cost

Potential economic value, continued

. Contingency table indicating the costs and losses accrued by the use of weather forecasts, depending on forecast and observed events.

Observation	Forecast/action: Yes	Forecast/action: No	
Yes	Hit (h) Mitigated loss ($C + L_u$)	Miss (m) Loss ($L = L_p + L_u$)	$h + m = \bar{o}$
No	False Alarm (f) Cost (C)	Correct rejection (c) No cost (N)	$f + c = 1 - \bar{o}$

Suppose we have the contingency table of forecast outcomes, [*h, m, f, c*].

Then we can calculate the expected value of the expenses from a forecast, from climatology, from a perfect forecast.

$$E_{forecast} = fC + h(C + L_u) + m(L_p + L_u)$$

$$E_{climate} = Min\left[\bar{o}(L_p + L_u),\ C + \bar{o}L_u\right] = \bar{o}L_u + Min\left[\bar{o}L_p,\ C\right]$$

$$E_{perfect} = \bar{o}(C + L_u)$$

$$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}} = \frac{Min\left[\bar{o}L_p,\ C\right] - (h + f)C - mL_p}{Min\left[\bar{o}L_p,\ C\right] - \bar{o}C}$$

Note that value will vary with C, L_p, L_u;

Different users with different protection costs may experience a different value from the forecast system.

From ROC to potential economic value

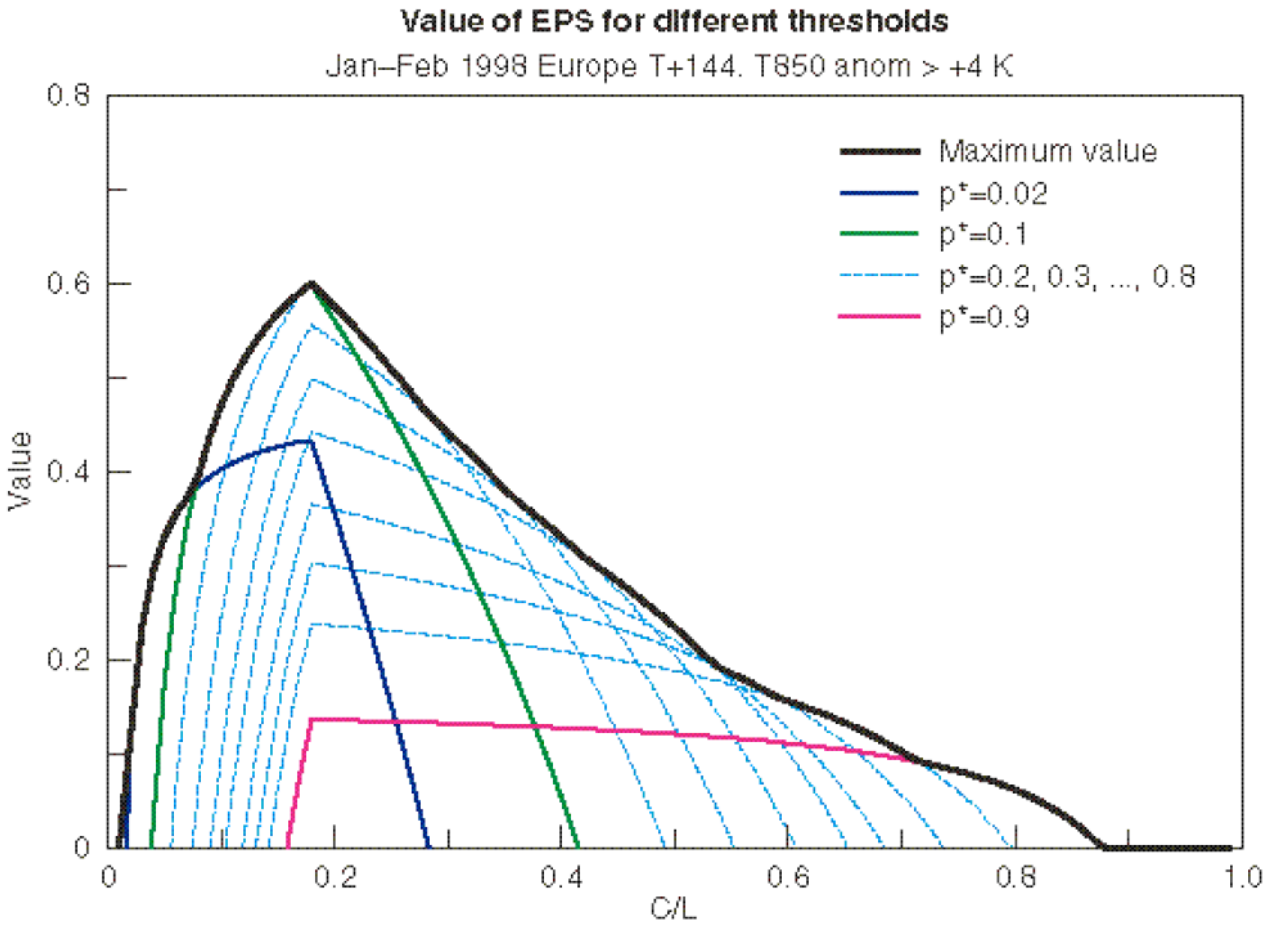
$$HR = \frac{h}{\bar{o}} \qquad FAR = \frac{f}{1-\bar{o}} \qquad m = \bar{o} - HR\,\bar{o}$$

$$V = \frac{Min\left[\bar{o}, C/L_p\right] - (h+f)C/L_p - m}{Min\left[\bar{o}, C/L_p\right] - or}$$

$$= \frac{Min\left[\bar{o}, C/L_p\right] - \left(C/L_p\right)FAR\left(1-\bar{o}\right) + HR\,\bar{o}\left(1 - C/L_p\right) - \bar{o}}{Min\left[\bar{o}, C/L_p\right] - \bar{o}r}$$

Value is now seen to be related to FAR and HR, the components of the ROC curve.

Economic value curve example



The red curve is from the ROC data for the member defining the 90th percentile of the ensemble distribution. Green curve is for the 10th percentile. Overall economic value is the maximum (use whatever member for decision threshold that provides the best economic value).

While admirable for framing verification in terms more relevant to the forecast user, the economic value calculations as presented here do not take into account other factors such as risk-aversion, or more complex decisions other than protect/don't.

Forecast skill often overestimated!

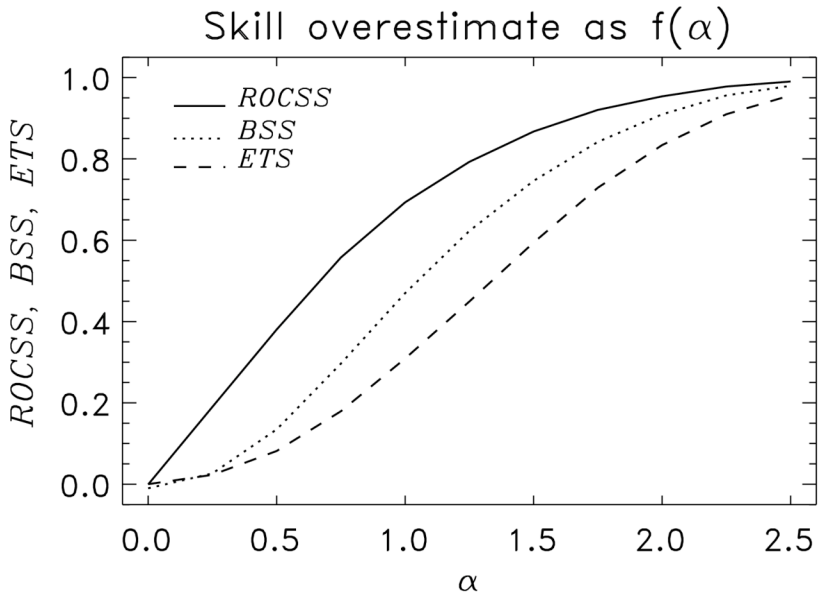
- Suppose you have a sample of forecasts from two islands, and each island has different climatology.

- Weather forecasts impossible on both islands.

- Simulate "forecast" with an ensemble of draws from climatology

- Island 1: F ~ N(α,1). Island 2: F ~ N(-α,1)

- Calculate ROCSS, BSS, ETS in normal way. Expect no skill.



As climatology of the two islands begins to differ, then "skill" increases though samples drawn from climatology.

These scores falsely attribute differences in samples' climatologies to skill of the forecast.

Samples must have the same climatological event frequency to avoid this.

reference: Hamill and Juras, *QJRMS*, Oct 2006

Other ensemble verification methods

- Bounding boxes (Judd et al., QJRMS, 2007; for similar idea, see Wilson et al., MWR, June 1999)
- Evaluation of linearity of forecast (Gilmour et al, JAS, 2001).
- Perturbation vs. error correlation (Toth et al., MWR, August 2003)
- Ignorance score (Roulston and Smith, MWR, June 2002)
- Discrimination diagram (Wilks text vol 2, 2006, p. 293)
- etc.

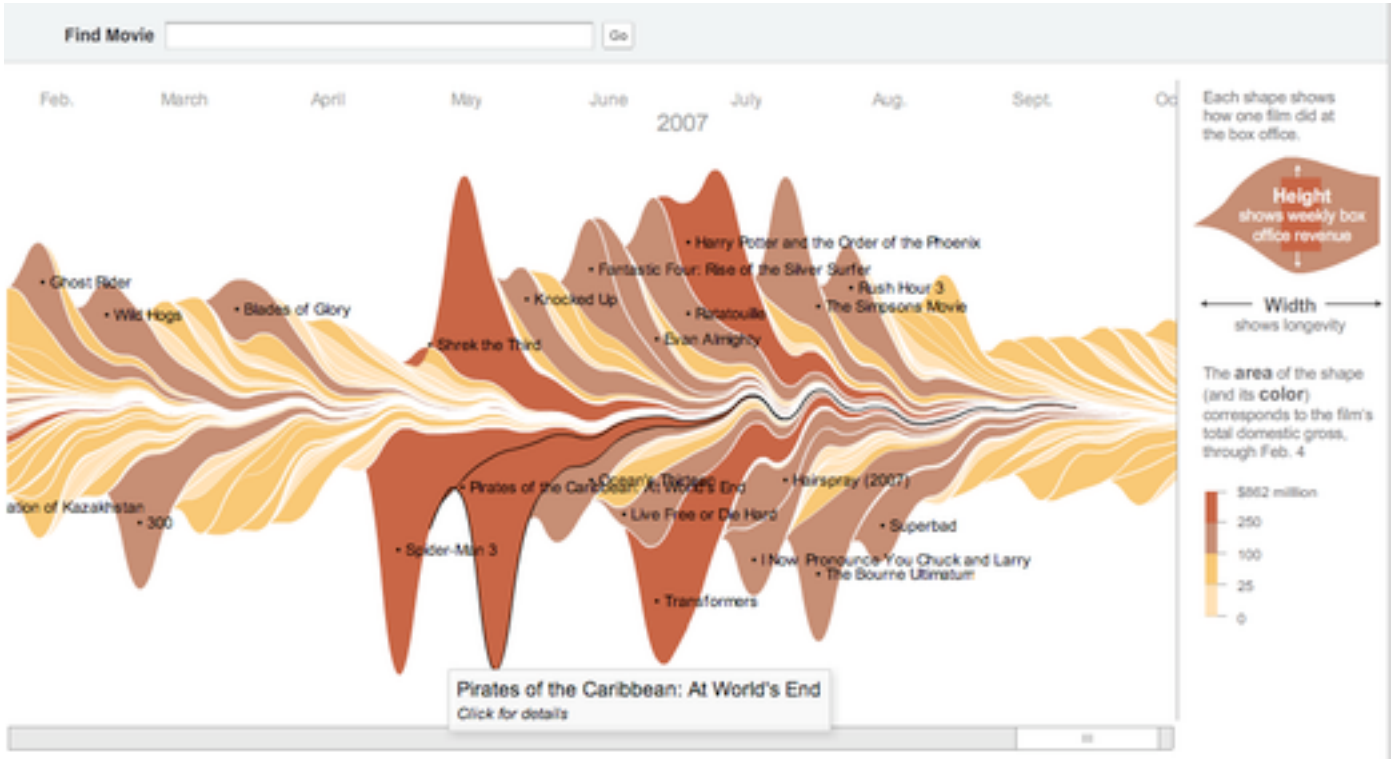
Visualization of ensemble forecast information

- Techiques primarily aimed at forecasters for interpretation of ensembles (convey the content of complex, high-information density data set in way that is maximally useful to forecaster)

  ⦚

- Techniques for conveying probabilistic information to the public effectively.

Example of dense information



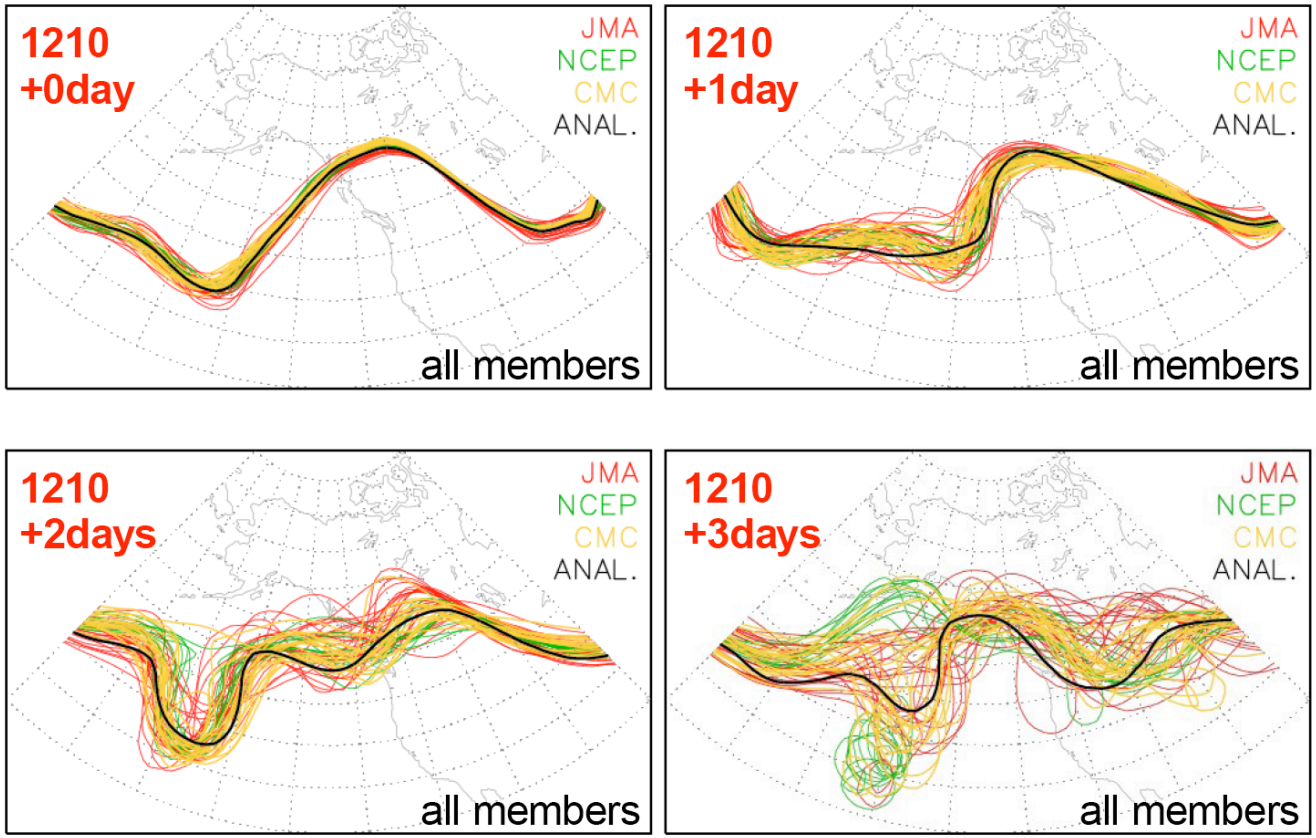
http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

Give the cognoscenti products that, once they understand them, will BLOW THEM AWAY.

Spaghetti diagrams

Example:

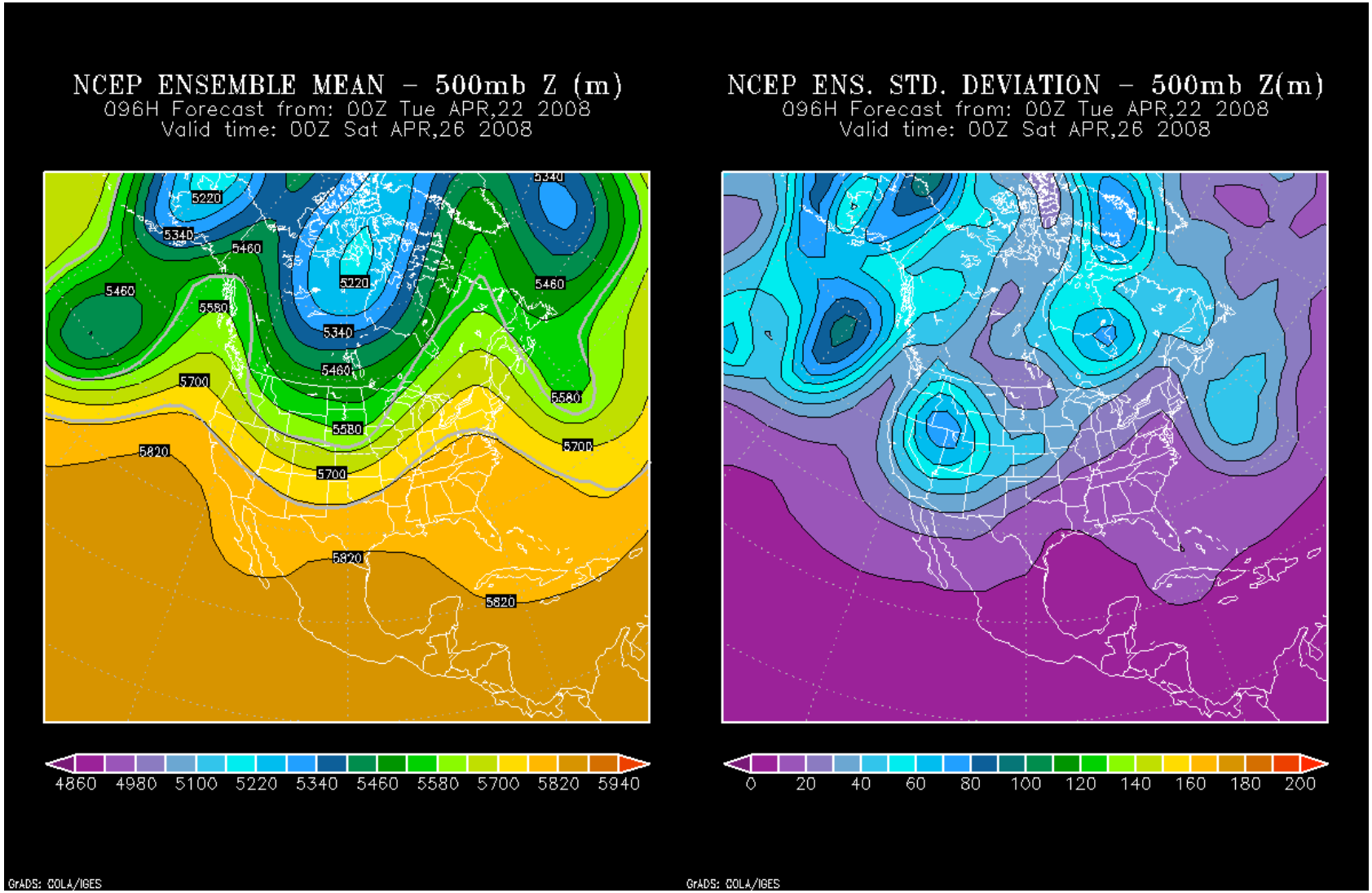
Z500 (5500m) Spaghetti Diagram initialized at 10th Dec. 2005



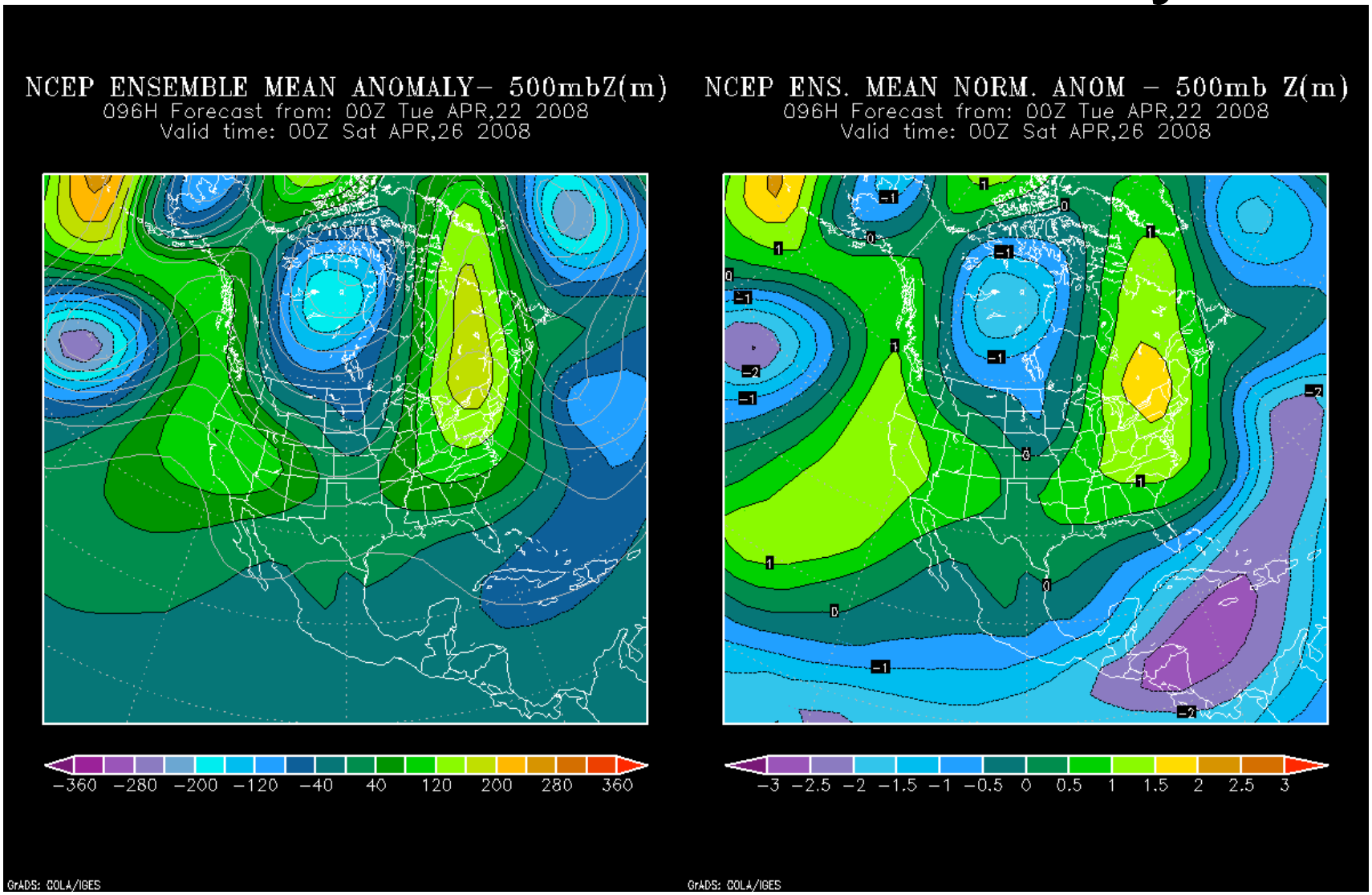
- A selected contour is plotted for each member.
- Advantage: provides a graphical representation of uncertainty.
- Disadvantage: representation can be misleading. In regions with weak gradients, will be large displacement of a member's line for a small change the forecast.

from Matsueda et al. presentation at 2nd International THORPEX symposium

Mean and standard deviation



Anomaly and normalized anomaly

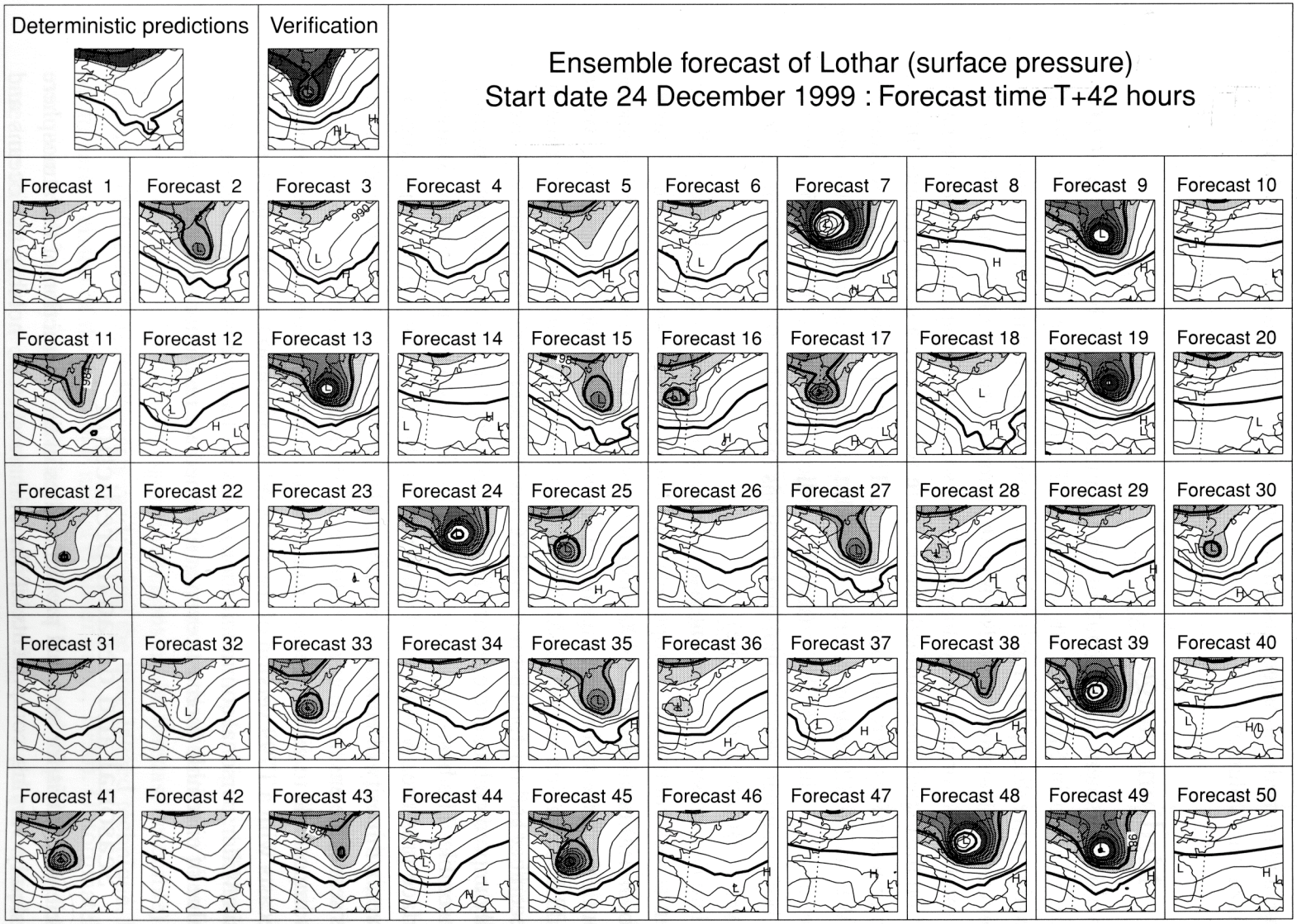


Stamp maps

Graphically shows each ensemble member:

Advantage: get to see the synoptic details of each member.

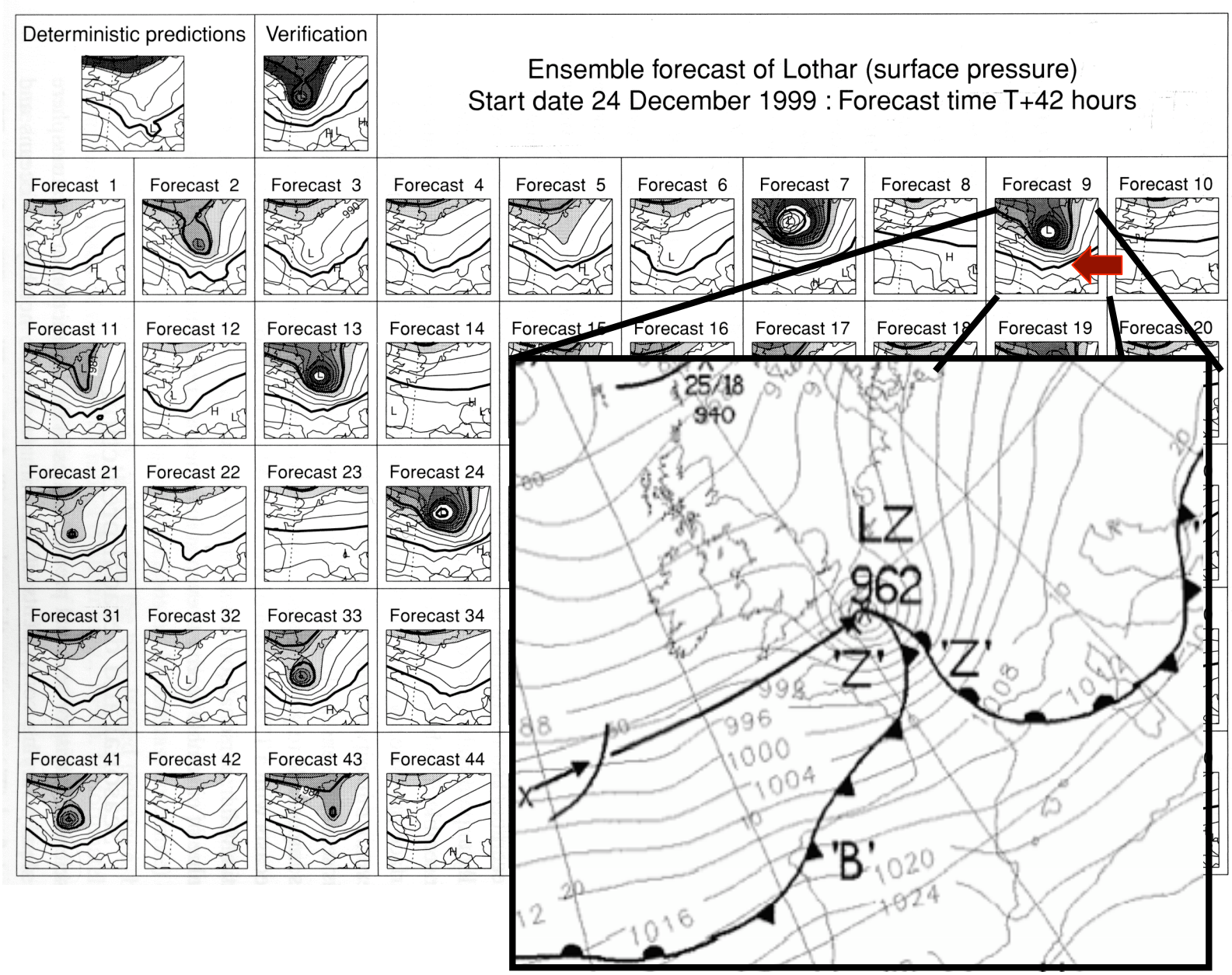
Disadvantage: With lots of members, small maps, and tough to show large areas / multiple fields at once.



from Tim Palmer's book chapter, 2006, in "*Predictability of Weather and Climate*".

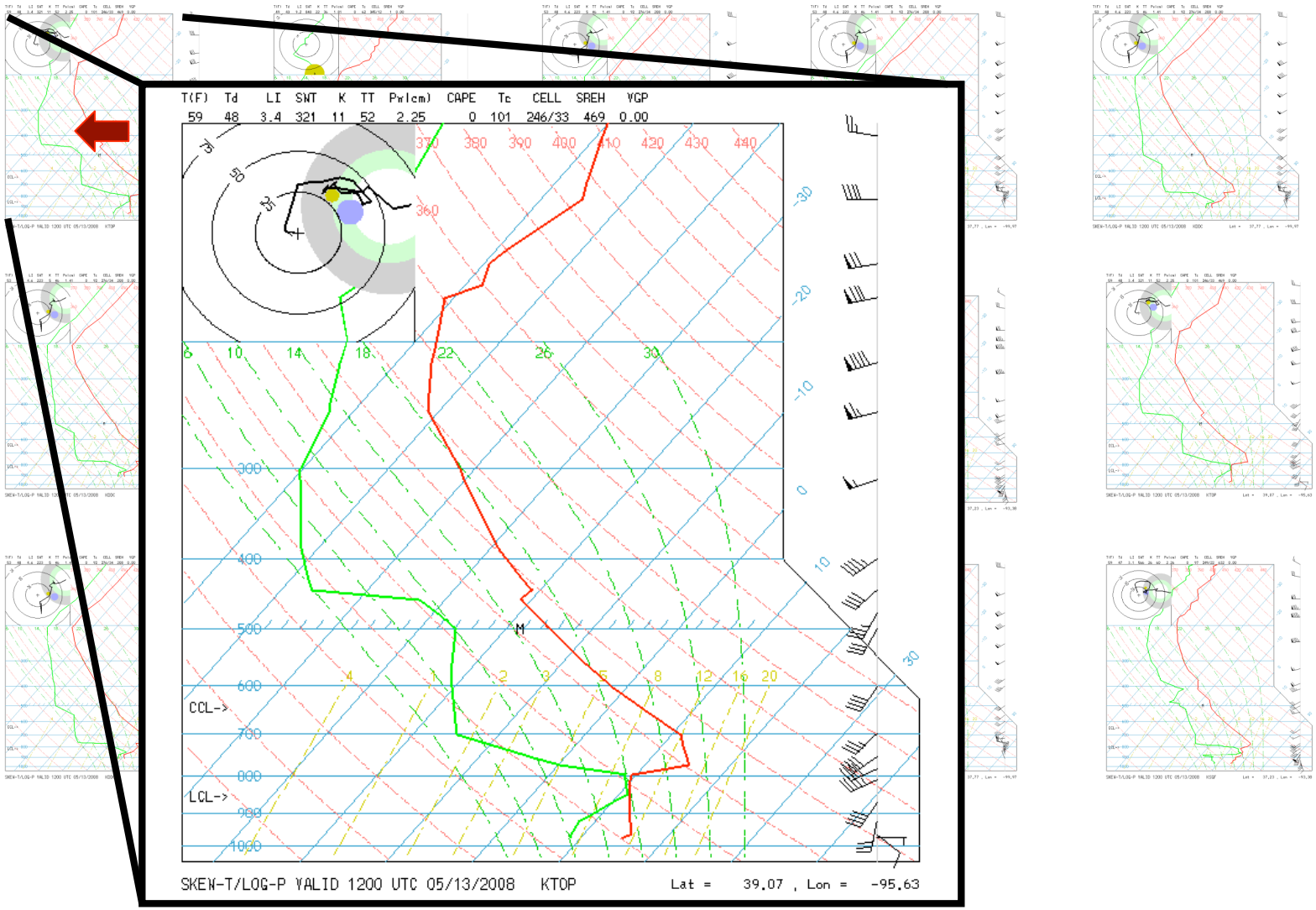
Stamp maps

Zoom capability with mouse over event



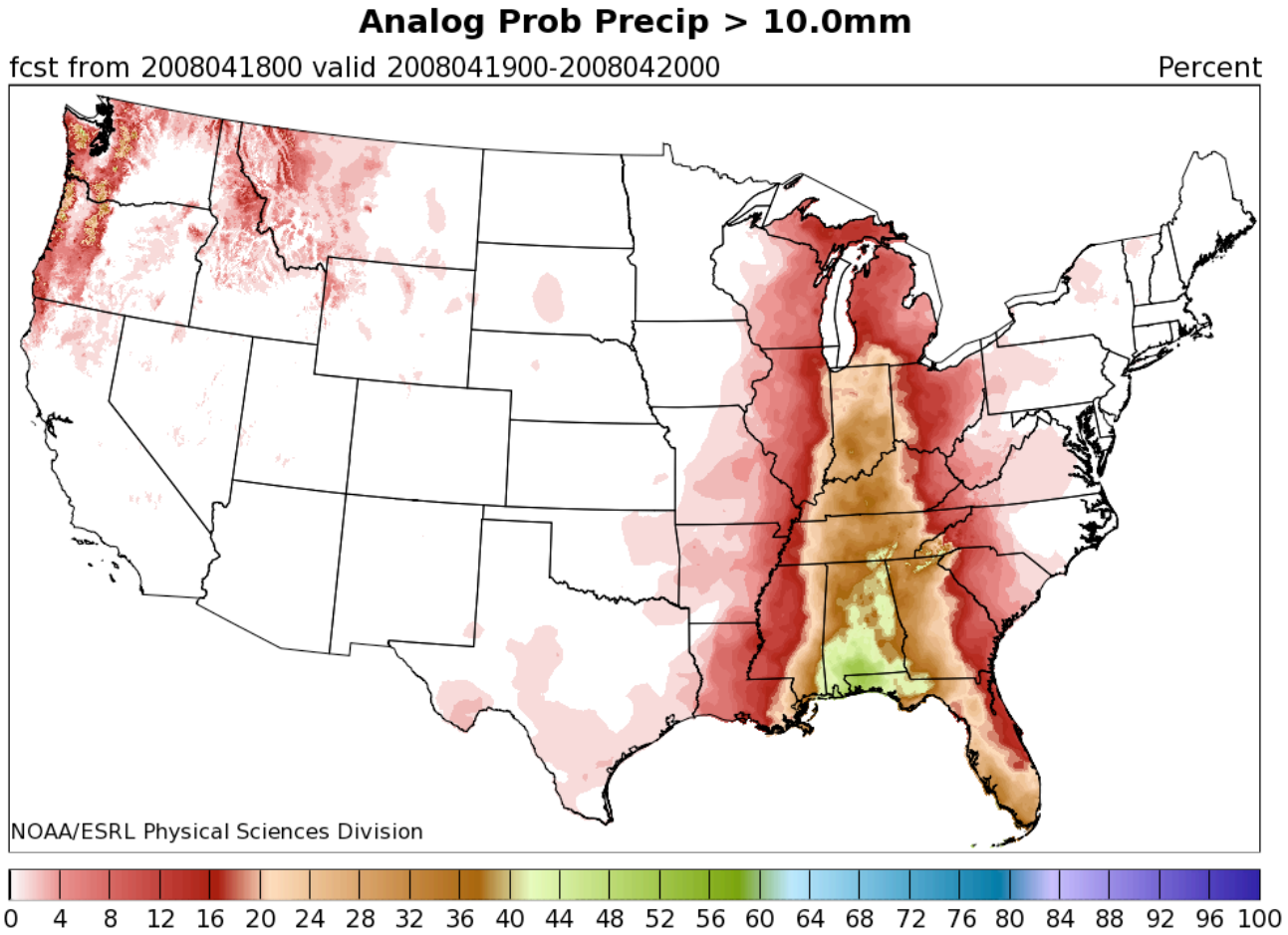
from Tim Palmer's book chapter, 2006, in "*Predictability of Weather and Climate*".

Stamp Skew-T's with mouse-over



Probability plots

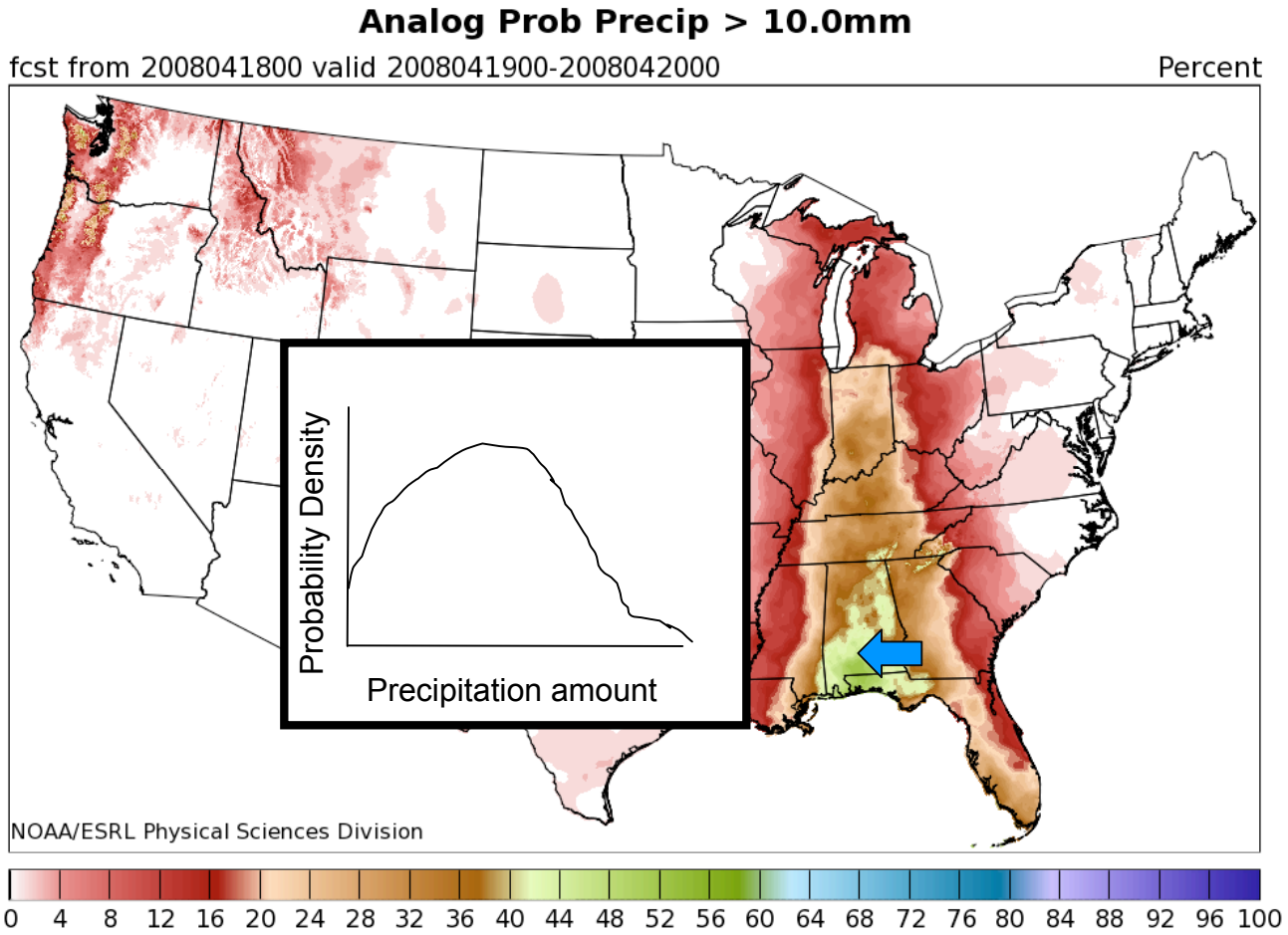
- Provides a graphical display of probabilities for a particular event, here for probability of greater than 10 mm rainfall in 24 h.
- Advantage: simple, relatively intuitive.
- Disadvantages: no sense of the meteorology involved, doesn't provide information on whole pdf.



from Hamill & Whitaker's analog reforecast technique web page, www.cdc.noaa.gov/reforecast/narr.

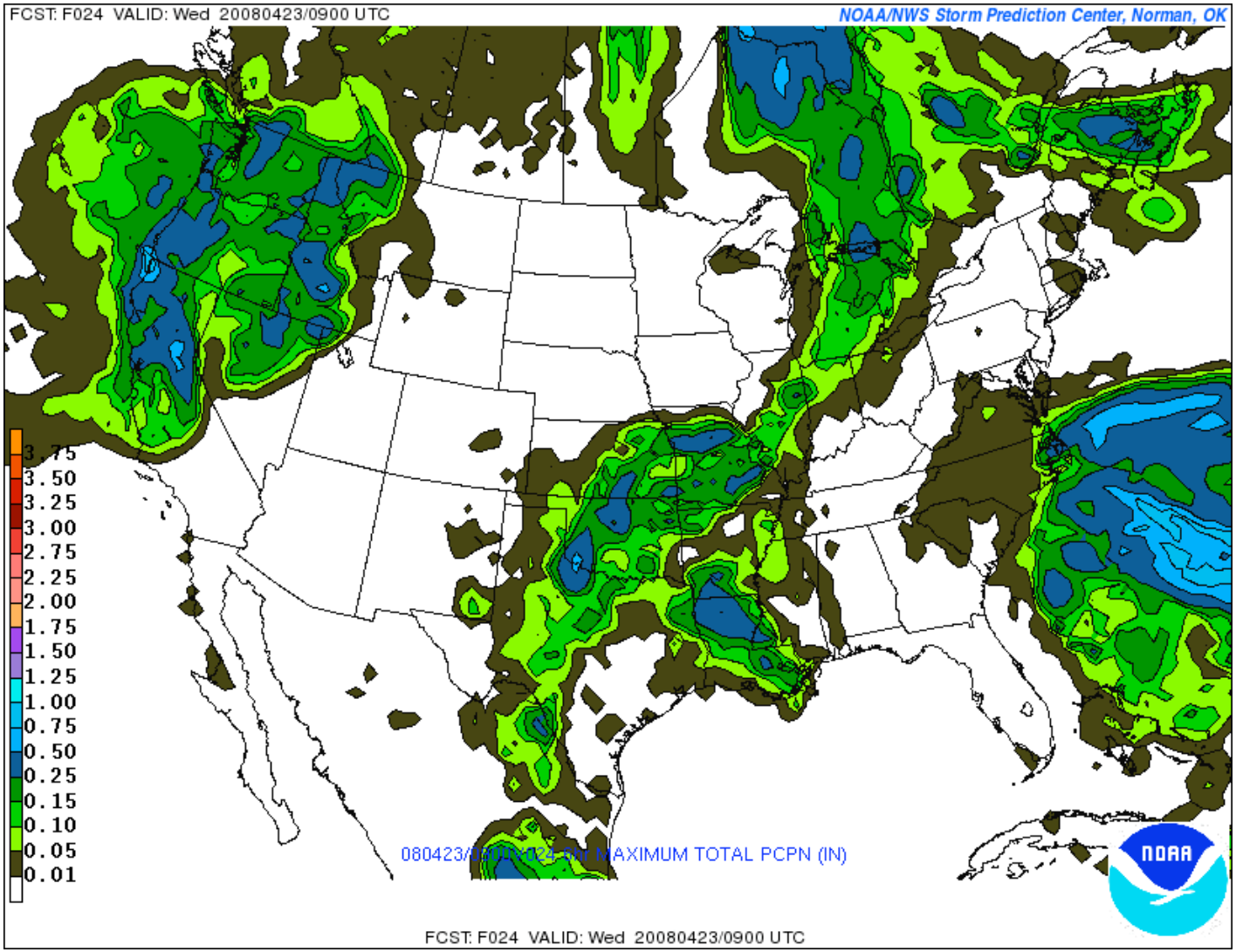
Probability plots

- With mouse-over event capability

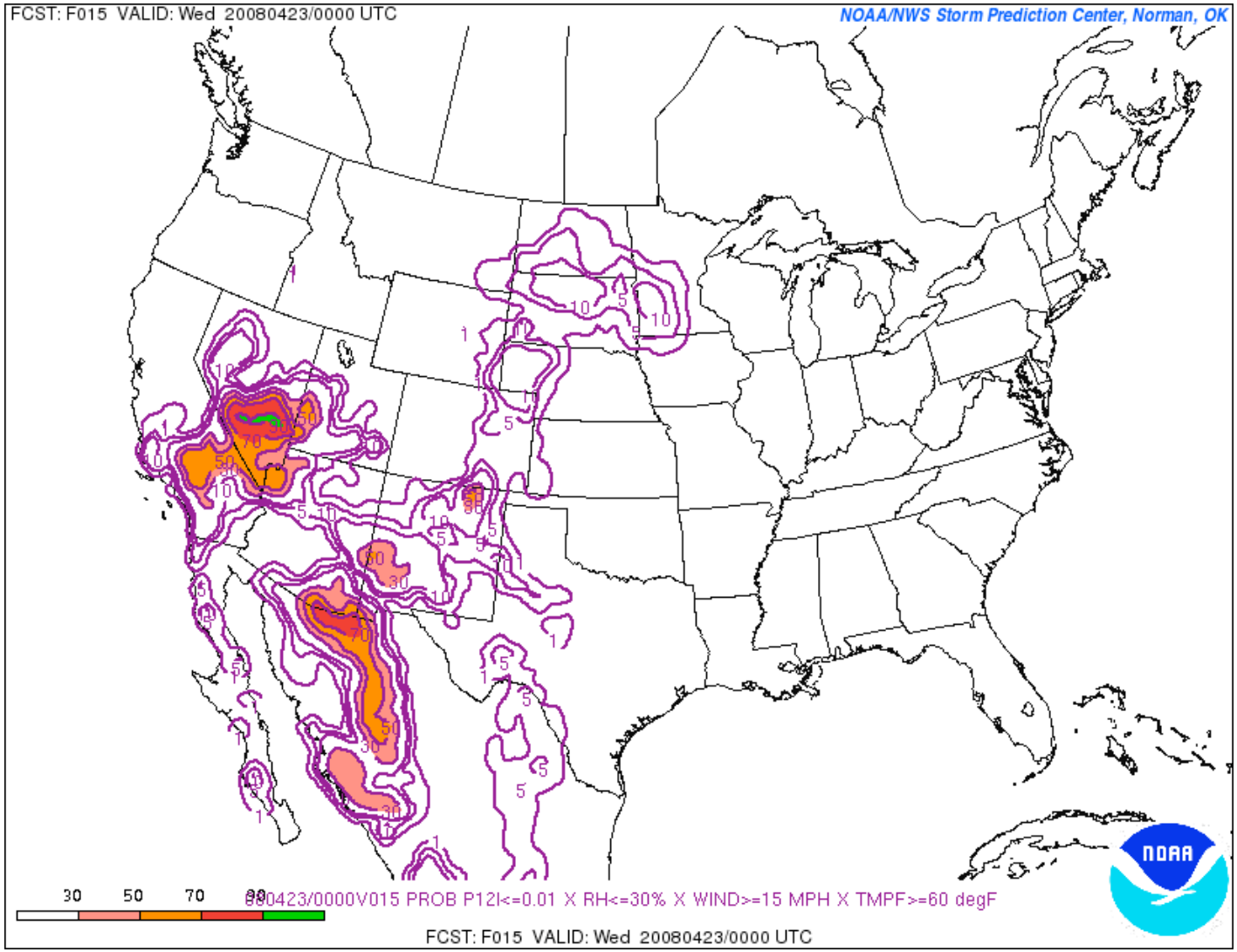


from Hamill & Whitaker's analog reforecast technique web page, www.cdc.noaa.gov/reforecast/narr.

Maximum 6-hourly total precipitation from all members

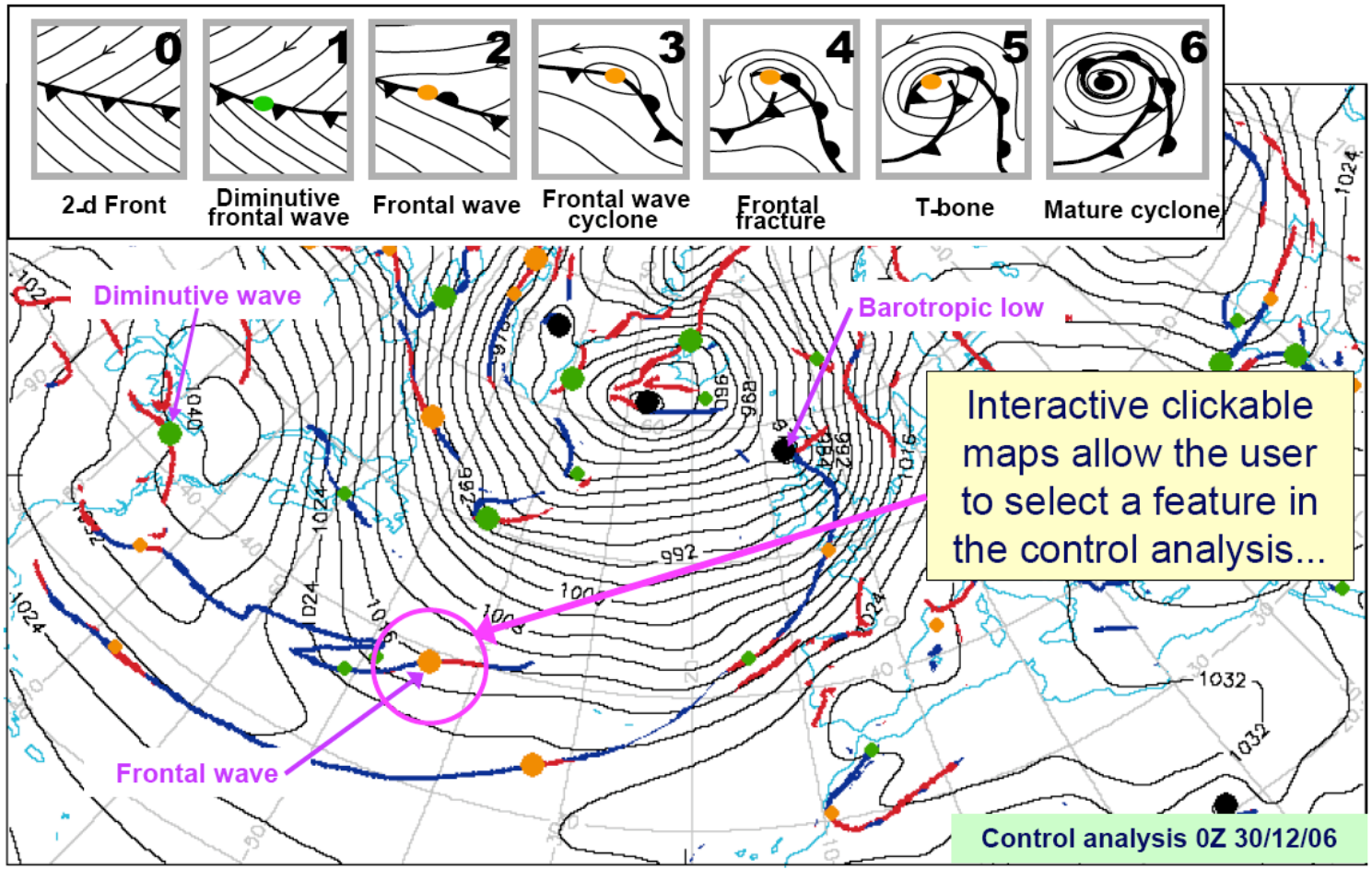


Joint probability of 12-hourly precip < 0.01 inches (~ .25 mm) and RH < 30% and wind speed > 15 mph (6.6 ms^{-1}) and T_{2m} > 60F (15.5 C)



here, useful for fire weather

Cyclone database & New Year's Eve storm

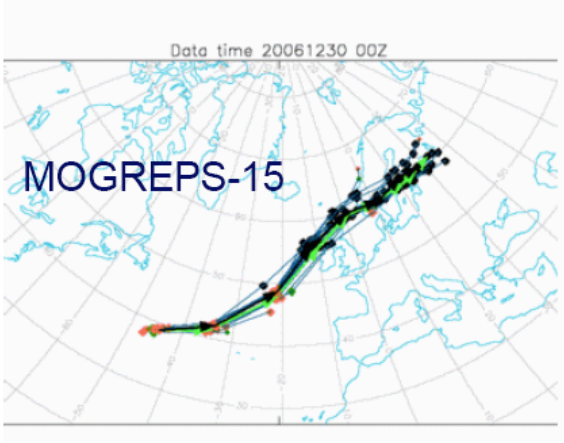


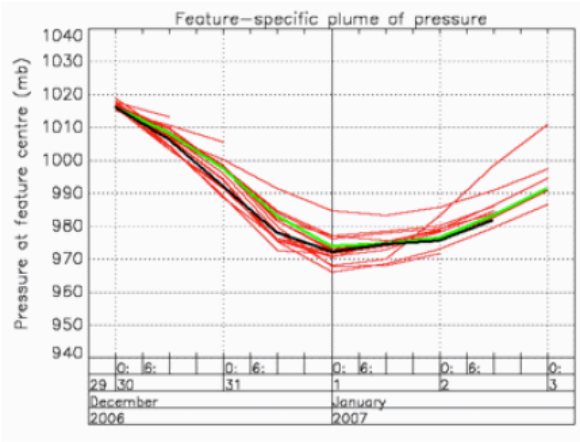
- Tracking scheme uses a combination of forward and backward tracking. It uses extrapolation and 500hPa steering wind to estimate positions, and matches features based on separation distance, type and thickness

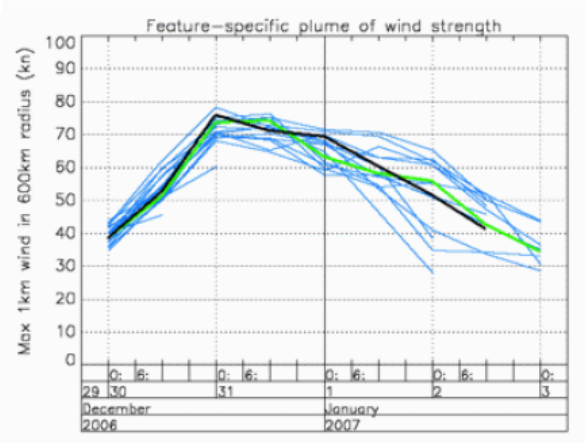
from Christine Johnson's presentation at Nov 2007 ECMWF workshop on ensemble prediction

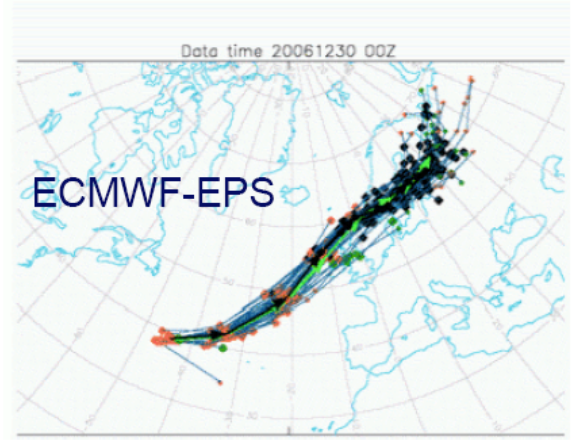
Cyclone database: 31/12/2006 example

- Clicking on a feature brings up feature-specific tracks from each ensemble member and matching plumes of intensity measures to identify the potential for high-impact weather

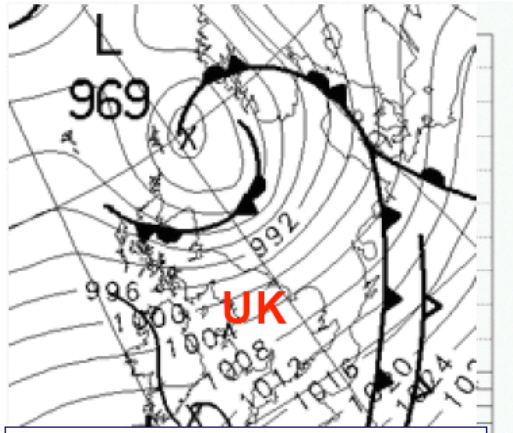








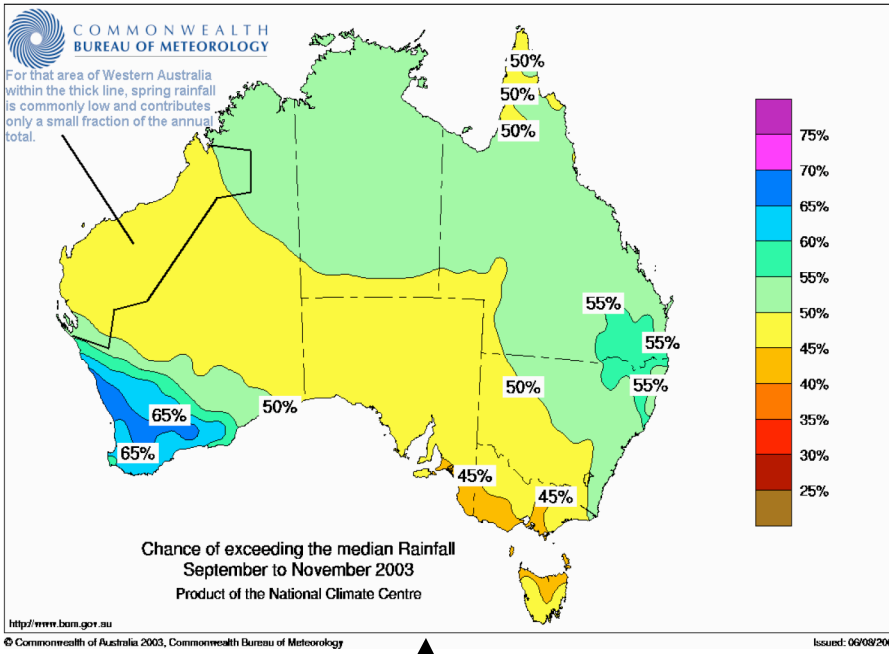
This storm tracked across Scotland, with gusts up to 100mph, leading to the high-profile cancellation of New Year's Eve celebrations and loss of power to 1000s of homes

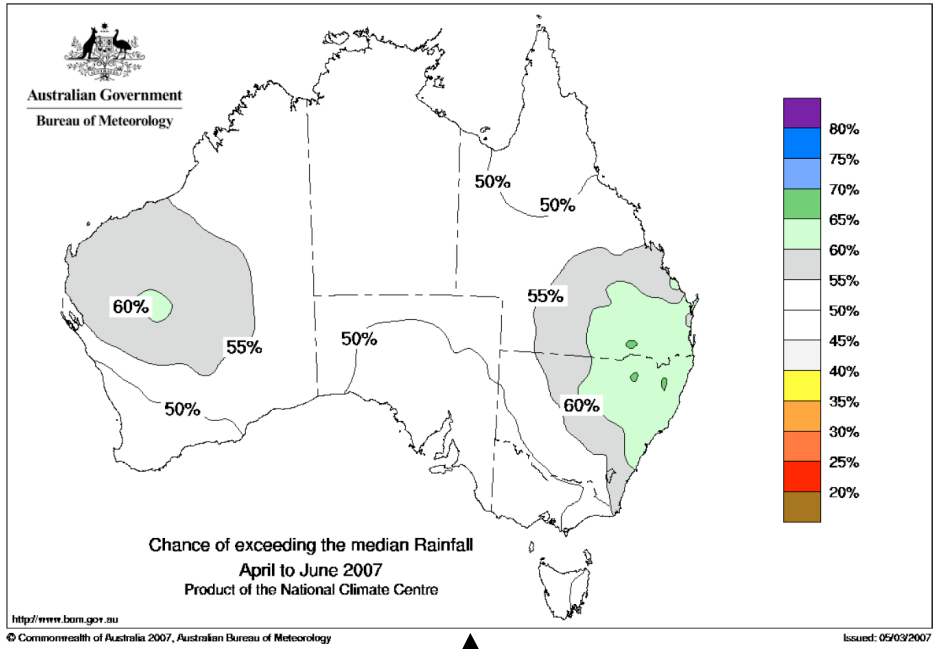


Analysis 00Z 01/01/2007

from Christine Johnson's presentation at Nov 2007 ECMWF workshop on ensemble prediction

Use and misuse of colors



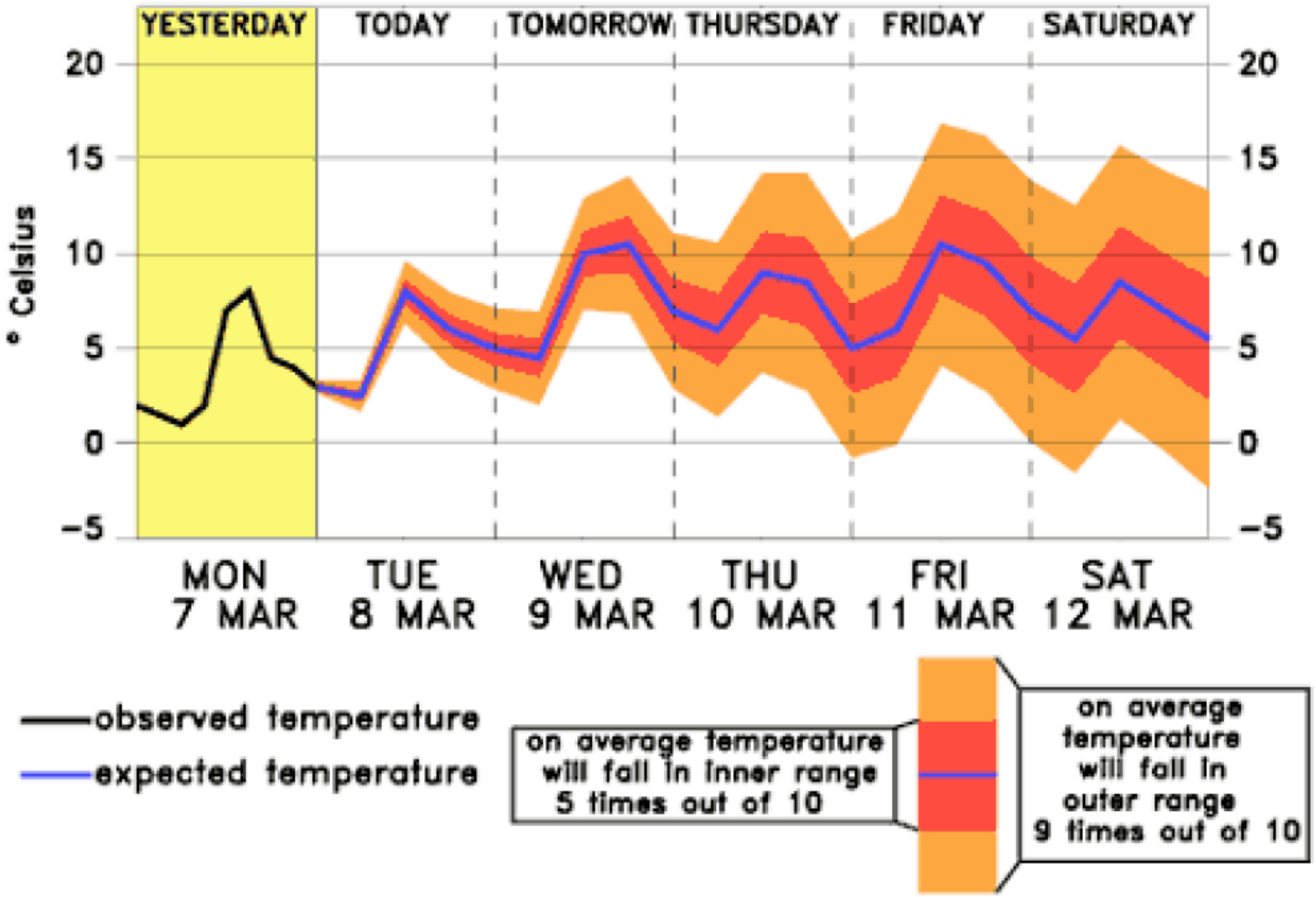


Bold colors for near 50% forecasts provide misleading sense of significance of small differences.

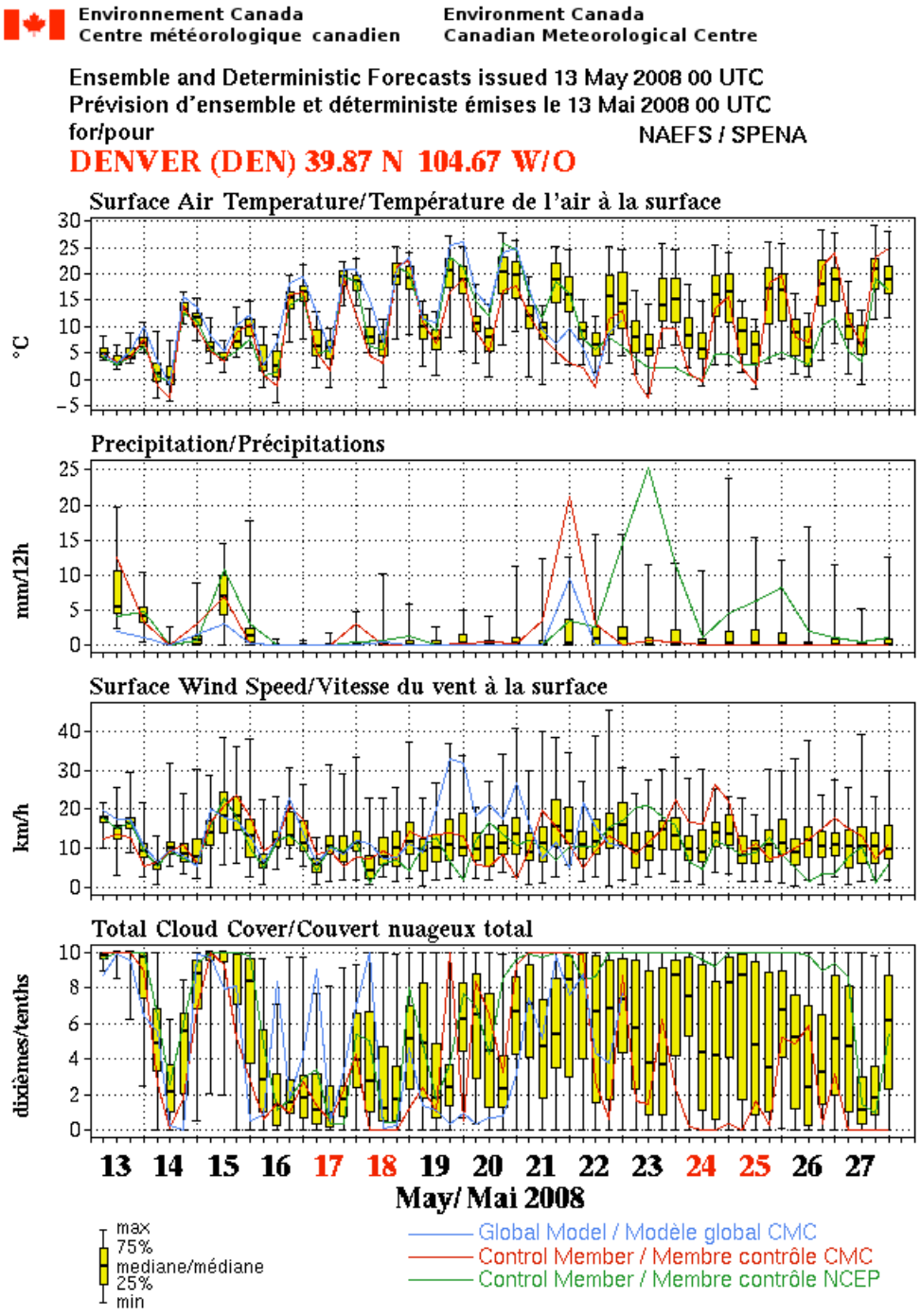
Better

from WMO/TD-1422, *Guidelines on Communicating Forecast Uncertainty*

Fan charts



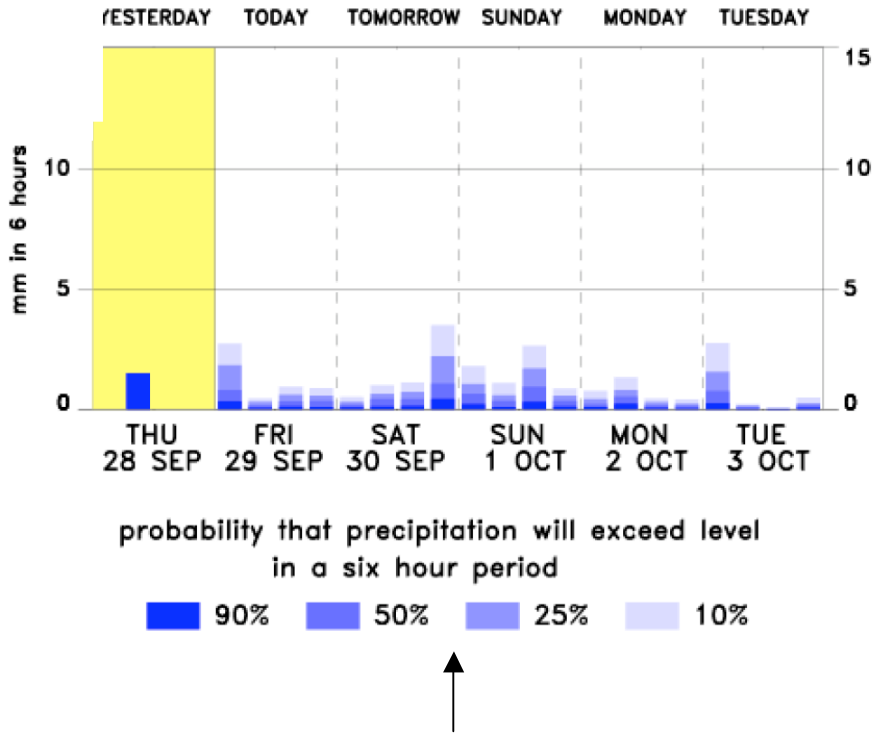
from Ken Mylne (Met Office) presentation to NWS NFUSE group



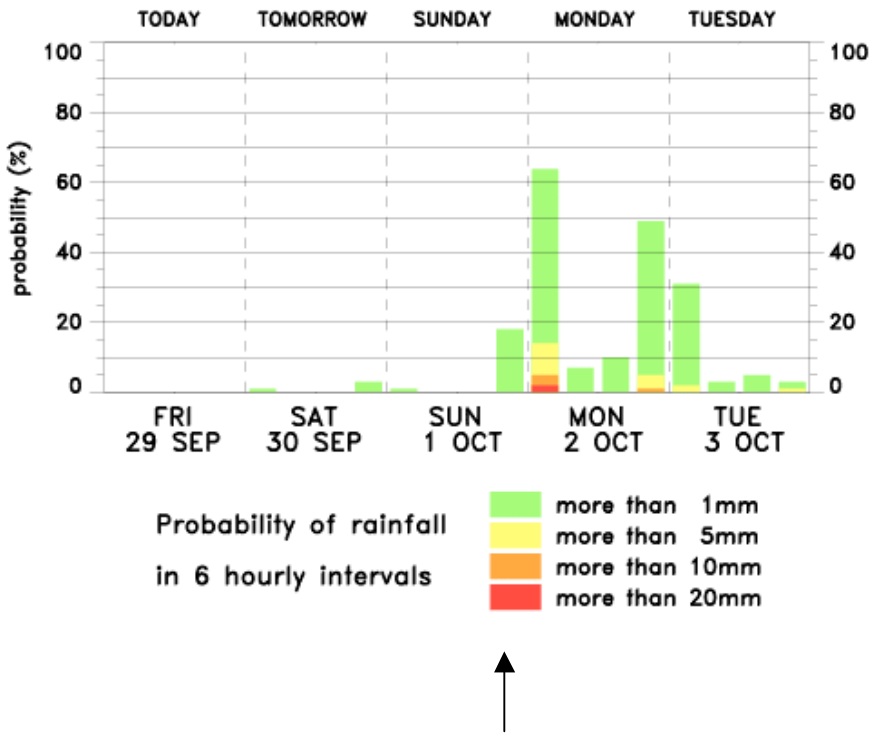
EPSgrams from RPN Canada

UK Met Office

user-preferred charts for precipitation



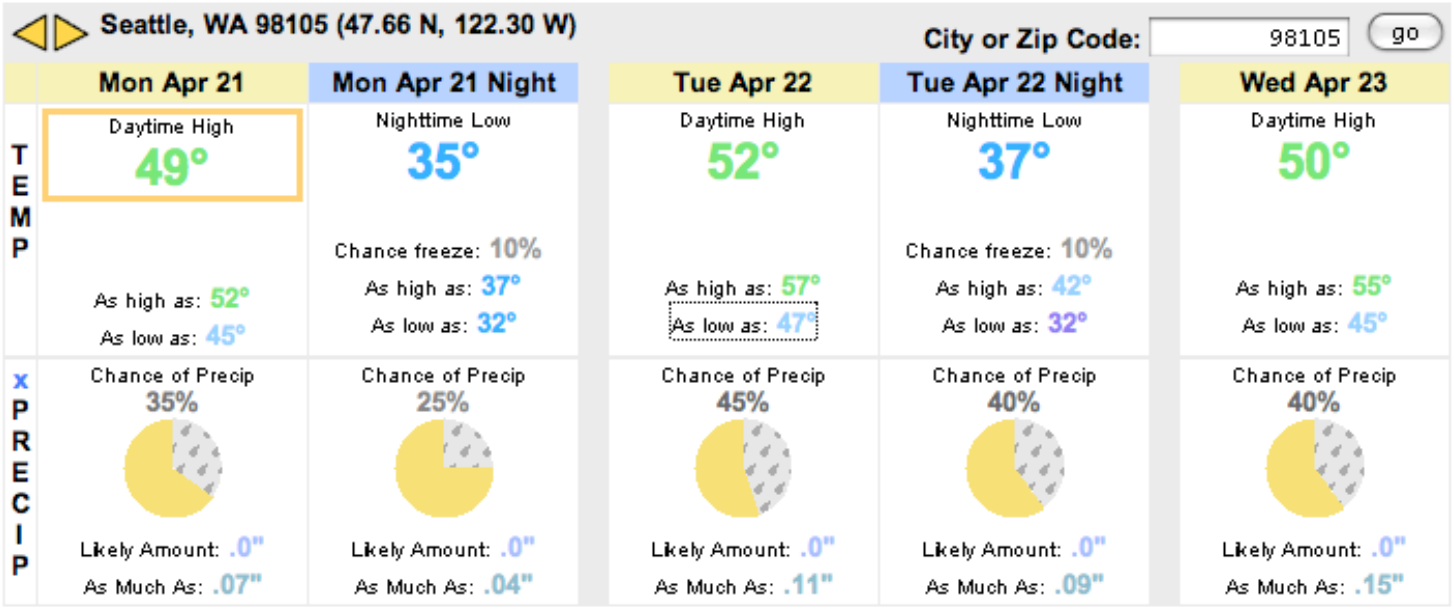
plots quantiles of the forecast pdf



plots exceedance probabilities

from Ken Mylne (Met Office) presentation to NWS NFUSE group

U. Washington's "Probcast"



Seattle, WA 98105 (47.66 N, 122.30 W)

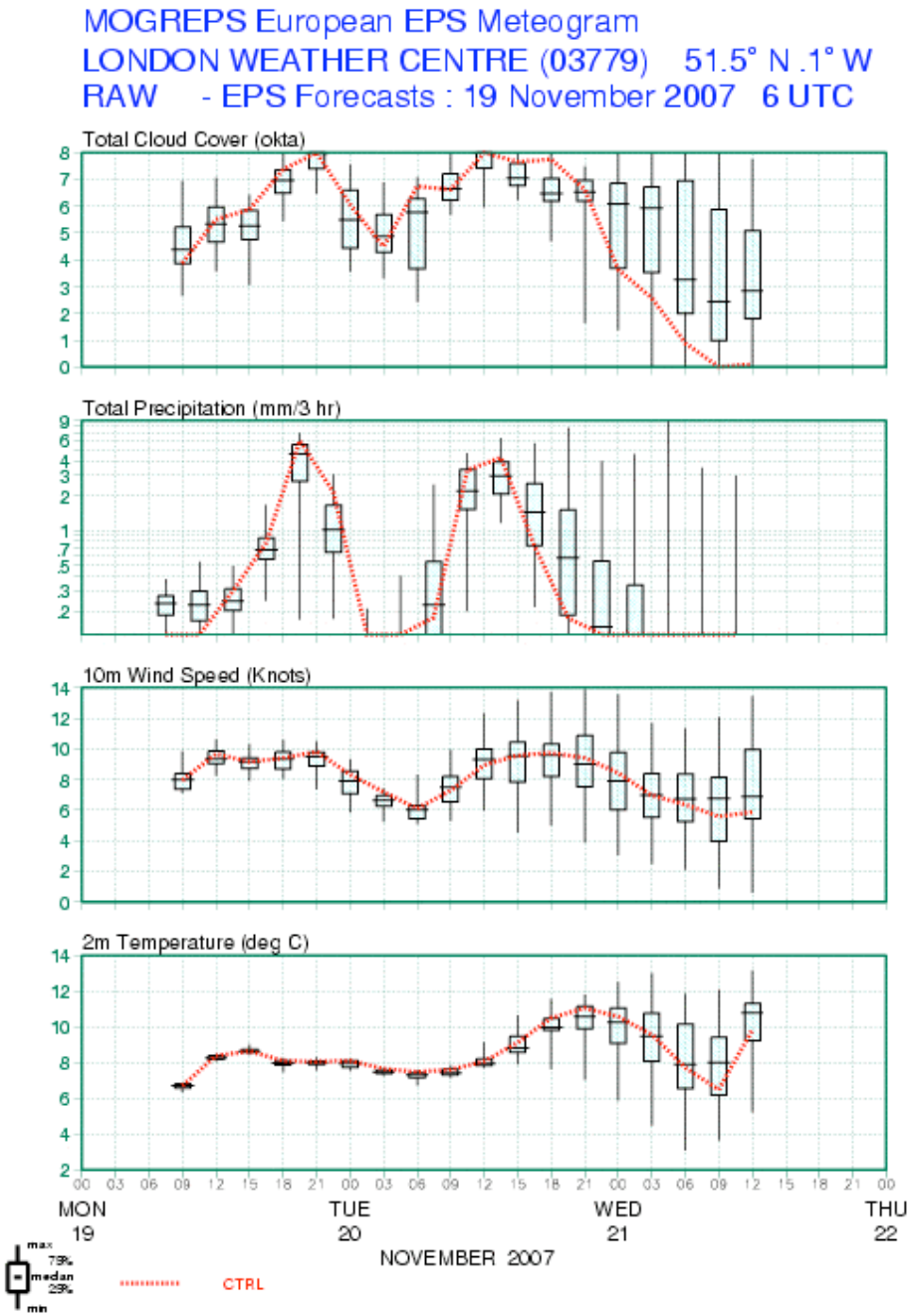
City or Zip Code: 98105 go

	Mon Apr 21	Mon Apr 21 Night	Tue Apr 22	Tue Apr 22 Night	Wed Apr 23
TEMP	Daytime High 49°	Nighttime Low 35°	Daytime High 52°	Nighttime Low 37°	Daytime High 50°
	As high as: 52°	Chance freeze: 10%	As high as: 57°	Chance freeze: 10%	As high as: 55°
	As low as: 45°	As high as: 37°	As low as: 47°	As high as: 42°	As low as: 45°
		As low as: 32°		As low as: 32°	
XPRECIP	Chance of Precip 35%	Chance of Precip 25%	Chance of Precip 45%	Chance of Precip 40%	Chance of Precip 40%
	Likely Amount: .0"	Likely Amount: .0"	Likely Amount: .0"	Likely Amount: .0"	Likely Amount: .0"
	As Much As: .07"	As Much As: .04"	As Much As: .11"	As Much As: .09"	As Much As: .15"

http://probcast.washington.edu

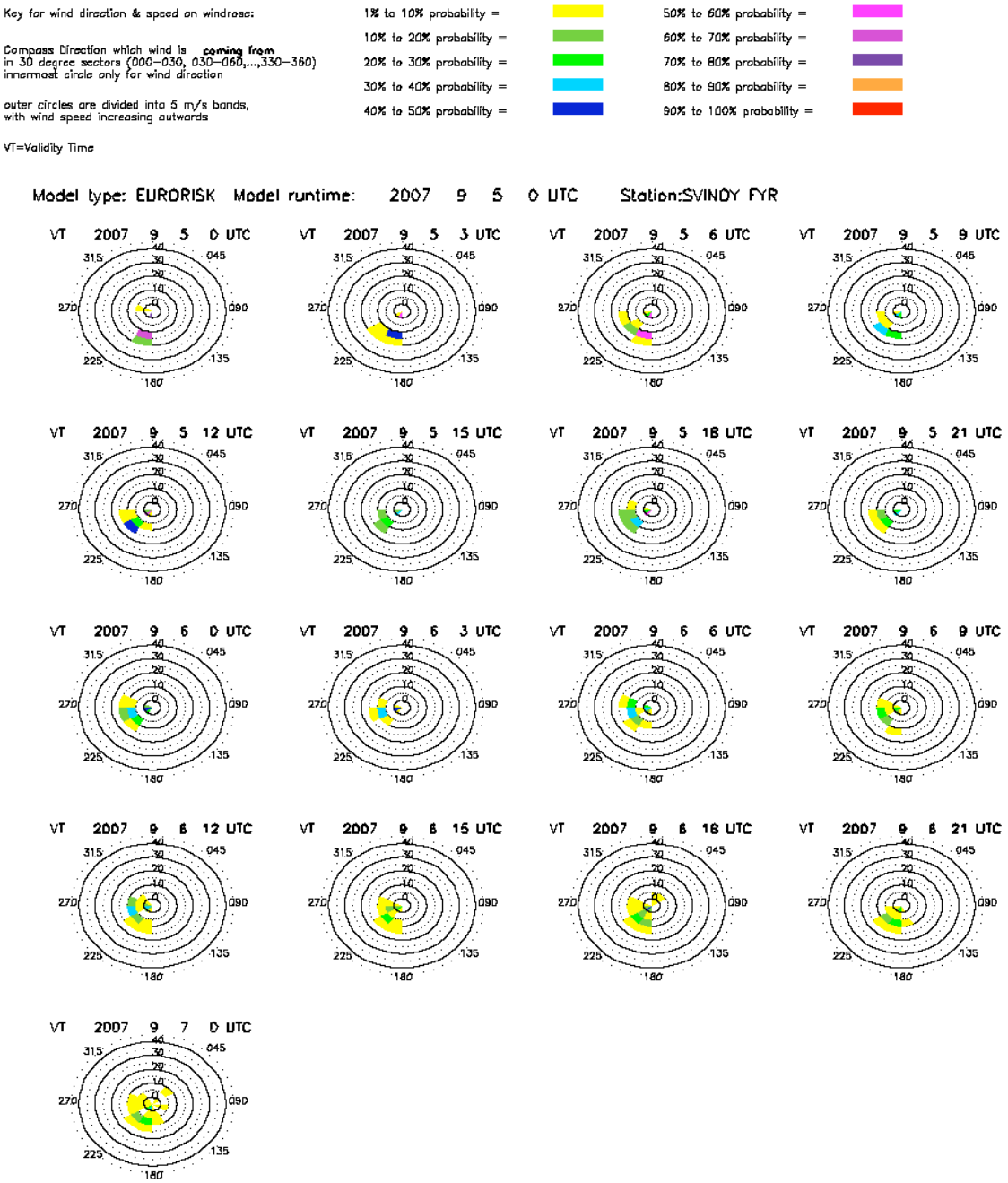
Meteograms

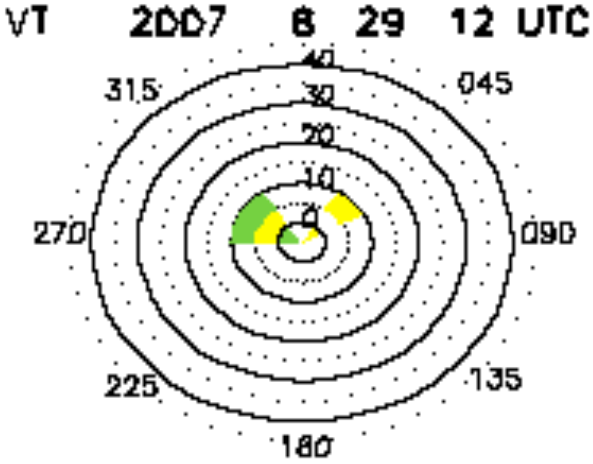
- original design by ECMWF
- widely used by ensemble forecasters
- min, max, 80th, 20th percentiles, plus median conveyed through “box and whiskers”



from Ken Mylne (Met Office) presentation to NWS NFUSE group.

Wind roses – probabilities of speed and direction





Verbal descriptions of uncertainty – the IPCC scale

The IPCC have proposed a likelihood scale for communication of climate change predictions:

Virtually Certain	> 99% probability
Very Likely	> 90% probability
Likely	> 66% probability
About as likely as not	33% to 66% probability
Unlikely	< 33% probability
Very Unlikely	< 10% probability
Exceptionally Unlikely	< 1% probability

from Ken Mylne (Met Office) presentation to NWS NFUSE group

Verbal descriptions of uncertainty – an alternative scale

An alternative scale proposed for general use by WMO

Extremely Likely	> 99% probability
Very Likely	90-99% probability
Likely	70-89% probability
Probable – more likely than not	55-69% probability
Equally likely as not	45-54% probability
Possible – less likely than not	30-44% probability
Unlikely	10-29% probability
Very Unlikely	1-9% probability
Extremely Unlikely	< 1% probability

from Ken Mylne (Met Office) presentation to NWS NFUSE group

Good resource for how to present complex information